

The application of genomic technologies to cancer and companion diagnostics.

James Ettore Hadfield BSc

PhD by Publication

University of East Anglia

School of Biological Sciences

September 2014

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

This thesis describes work undertaken by the author between 1996 and 2014. Genomics is the study of the genome, although it is also often used as a catchall phrase and applied to the transcriptome (study of RNAs) and methylome (study of DNA methylation). As cancer is a disease of the genome the rapid advances in genomic technology, specifically microarrays and next generation sequencing, are creating a wave of change in our understanding of its molecular pathology. Molecular pathology and personalised medicine are being driven by discoveries in genomics, and genomics is being driven by the development of faster, better and cheaper genome sequencing. The next decade is likely to see significant changes in the way cancer is managed for individual cancer patients as next generation sequencing enters the clinic.

In chapter 3 I discuss how *ERBB2* amplification testing for breast cancer is currently dominated by immunohistochemistry (a single-gene test); and present the development, by the author, of a semi-quantitative PCR test for *ERBB2* amplification. I also show that estimating *ERBB2* amplification from microarray copy-number analysis of the genome is possible. In chapter 4 I present a review of microarray comparison studies, and outline the case for careful and considered comparison of technologies when selecting a platform for use in a research study. Similar, indeed more stringent, care needs to be applied when selecting a platform for use in a clinical test. In chapter 5 I present co-authored work on the development of amplicon and exome methods for the detection and quantitation of somatic mutations in circulating tumour DNA, and demonstrate the impact this can have in understanding tumour heterogeneity and evolution during treatment. I also demonstrate how next-generation sequencing technologies may allow multiple genetic abnormalities to be analysed in a single test, and in low cellularity tumours and/or heterogenous cancers.

Keywords: Genome, exome, transcriptome, amplicon, next-generation sequencing, differential gene expression, RNA-seq, ChIP-seq, microarray, *ERBB2*, companion diagnostic.

Table of Contents

ABSTRACT	3
TABLE OF CONTENTS	5
LIST OF FIGURES	9
LIST OF TABLES	10
LIST OF ACCOMPANYING MATERIAL	11
ACKNOWLEDGMENTS	13
CHAPTER 1: INTRODUCTION	15
GENOMICS TECHNOLOGY	17
DNA SEQUENCING	17
NEXT-GENERATION SEQUENCING	18
THE GENOME	19
ANALYSIS OF THE HUMAN GENOME ALLOWS ESTIMATION OF THE NUMBER OF GENES, MANY OF WHICH DO NOT MAKE PROTEINS	19
GENES ARE COMPRISED OF EXONS, INTRONS AND REGULATORY SEQUENCES	20
THE HUMAN GENOME PROJECT	20
CANCER	21
CANCER IS A HETEROGENEOUS SET OF DISEASES	21
BREAST CANCER IS THE SECOND MOST COMMON CAUSE OF CANCER MORTALITY IN WOMEN	22
THE CANCER GENOME	22
DIFFERENT MUTATIONAL MECHANISMS UNDERLIE THE CAUSES OF CANCER	27
TUMOUR SUPPRESSOR GENES AND ONCOGENES DRIVE CANCER	27
SUMMARY	28
CHAPTER 2: METHODS PRESENTED, EXPERIMENTAL DESIGN AND QUALITY CONTROL	31
INTRODUCTION	33
METHODS PRESENTED	33
IMMUNOHISTOCHEMISTRY	33

REAL-TIME QUANTITATIVE PCR	34
METHODS TO DETECT AMPLICONS GENERATED BY QPCR	34
STANDARDISING QPCR EXPERIMENTS	41
QPCR AS A DIAGNOSTIC TOOL	41
MICROARRAY	42
MICROARRAY ANALYSIS OF RNA: DIFFERENTIAL GENE EXPRESSION (DGE) AND MICRO-RNA	42
MICROARRAY ANALYSIS OF COPY-NUMBER VARIATION (CNV)	47
MICROARRAY AS A DIAGNOSTIC TOOL	47
NEXT-GENERATION SEQUENCING	48
ILLUMINA SEQUENCING BY SYNTHESIS	48
RNA-SEQUENCING (RNA-SEQ)	48
CHROMATIN IMMUNOPRECIPITATION SEQUENCING (CHIP-SEQ)	53
EXOME-SEQUENCING (EXOME-SEQ)	53
EXPERIMENTAL DESIGN	53
THE IMPACT OF REPLICATION ON EXPERIMENTAL DESIGN	54
EXPERIMENTAL FACTORS AFFECTING REPLICATION AND EXPERIMENTAL DESIGN	55
QUALITY CONTROL	55
QUALITY CONTROL OF NUCLEIC ACIDS	55
QUALITY CONTROL IN NEXT-GENERATION SEQUENCING	56
CHAPTER 3: THE DEVELOPMENT OF COMPANION DIAGNOSTICS	61
INTRODUCTION	63
ERBB2:	63
MEASURING ERBB2:	64
MEASURING ERBB2 WITH IHC:	67
MEASURING ERBB2 WITH FISH:	67
MEASURING ERBB2 WITH MICROARRAYS:	68
MEASURING ERBB2 WITH END-POINT PCR:	68
DEVELOPMENT OF A DIFFERENTIAL-PCR METHOD FOR ERBB2 AMPLIFICATION STATUS	68
OPTIMISATION OF THE ERBB2 DIFFERENTIAL-PCR	68

OTHER PCR-BASED METHODS AND CITATION OF OUR DIFFERENTIAL-PCR TEST	69
THE IMPACT OF JENNINGS ET AL ON	70
COMPANION DIAGNOSTICS	73
CHAPTER 4: COMPARING THE ANALYTICAL VALIDITY OF PLATFORMS FOR GENOME AND GENE EXPRESSION STUDIES	77
INTRODUCTION:	79
ASSESSING THE QUALITY OF COMPARISONS:	79
BIAS IN COMPARISON STUDIES	80
A SEMI-SYSTEMATIC REVIEW OF MRNA MICROARRAY COMPARISON STUDIES	81
INCLUSION AND EXCLUSION CRITERIA	81
DATA EXTRACTION	81
RESULTS	82
THE MICROARRAY QUALITY CONTROL (MAQC) CONSORTIUM SETS THE STANDARD FOR, AND PROVIDES REFERENCE STANDARDS TO USE IN, COMPARISON STUDIES	82
GOLD-STANDARDS	87
COMPARISON PAPERS PUBLISHED BY THE AUTHOR	87
CURTIS ET AL 2009: THE PITFALLS OF PLATFORM COMPARISON: DNA COPY NUMBER ARRAY TECHNOLOGIES ASSESSED.	87
CONSIDERING BIAS IN THE DESIGN OF CURTIS ET AL 2009.	87
SELECTING CONTROL SAMPLES FOR CURTIS ET AL 2009.	88
RESULTS FROM CURTIS ET AL 2009.	88
GIT ET AL 2010: SYSTEMATIC COMPARISON OF MICROARRAY PROFILING, REAL-TIME PCR, AND NEXT-GENERATION SEQUENCING TECHNOLOGIES FOR MEASURING DIFFERENTIAL MICRORNA EXPRESSION.	91
GIT ET AL 2010 REQUIRED A LARGER AND MORE COMPLEX EXPERIMENTAL DESIGN.	91
CONTROLS AND VALIDATION USED IN GIT ET AL 2010.	92
CONCLUSIONS	92
CHAPTER 5: APPLICATION OF GENOMIC TECHNOLOGIES IN TRANSLATIONAL CANCER RESEARCH	101
INTRODUCTION	103
THE IMPACT OF NEXT-GENERATION SEQUENCING ON CANCER BIOLOGY	103
NGS CANCER STUDIES REVEAL EVOLUTIONARY MECHANISMS	104

ESTIMATING THE AMOUNT OF SEQUENCING REQUIRED:	104
CHOOSING BETWEEN AMPLICONS, EXOMES AND GENOMES	104
APPLYING NGS TECHNOLOGIES TO CIRCULATING TUMOUR DNA	108
CELL FREE DNA	108
DEVELOPMENT OF CTDNA AMPLICON SEQUENCING METHODS	108
THE CLINICAL UTILITY OF CTDNA AND TAM-SEQ	111
OTHER METHODS FOR ASSAYING CTDNA DO NOT COMPARE WELL TO TAM-SEQ	112
DEVELOPMENT OF CTDNA EXOME SEQUENCING METHODS	112
SENSITIVITY AND SPECIFICITY OF CTDNA ANALYSIS:	119
THE CLINICAL UTILITY OF CTDNA	120
CURRENT STATUS OF CLINICAL TESTING AND ADOPTION OF NGS ASSAYS	120
CHAPTER 6: DISCUSSION	123
ONCOGENES AND TUMOUR SUPPRESSOR GENES CAN BE ANALYSED TO DETERMINE THERAPY	125
PERSONALISED MEDICINE AND COMPANION DIAGNOSTICS ARE STILL IN THEIR INFANCY	126
PATIENT RESPONSE TO THERAPY IS HETEROGENOUS	126
PERSONALISED MEDICINE OFFERS SIGNIFICANT OPPORTUNITIES FOR TREATING CANCER	129
IMPROVED TESTING IS PART OF THE ANSWER	130
LIMITATIONS IN THE MOLECULAR ANALYSIS OF CANCER SAMPLES: NUCLEIC ACID QUALITY	130
LIMITATIONS IN THE MOLECULAR ANALYSIS OF CANCER SAMPLES: TUMOUR HETEROGENEITY AND STROMAL CONTAMINATION	139
UNDERSTANDING CANCER BIOLOGY IS VITAL	139
THE FUTURE FOR NGS IN CANCER GENOMICS AND COMPANION DIAGNOSTICS	140
SUMMARY	141
DEFINITIONS	143
GLOSSARY	145
REFERENCES	155
APPENDIX 1: LETTERS OF SUPPORT	173
APPENDIX 2: PUBLICATIONS SUBMITTED	209

List of figures

Fig 1.1: Cancer is caused by sequential mutations of specific oncogenes and tumour suppressor genes.

Fig 1.2: The hallmarks of cancer simplify the complexity of cancer biology by describing cancers driving pathways.

Fig 2.1: *ERBB2* amplification testing by IHC

Fig2.2: qPCR is a sensitive method for measuring nucleic acids

Fig 2.3: Two different qPCR detection methods

Fig 2.4: Methods for microarray manufacture

Fig 2.5: Microarrays for mRNA differential gene expression

Fig 2.6: Microarrays for CNV & LOH

Fig 2.7: Illumina next generation sequencing technology

Fig2.8: RNA input affects DGE sensitivity

Fig2.9: ChIP-seq QC by capillary electrophoresis and analysis of a genome browser track

Fig 3.1: Incidence of *ERBB2* amplification in Pancreatic cancer data from ICGC

Fig 3.2: Differential PCR can be used to quantify *ERBB2* amplification

Fig 3.3: List of FDA Companion Diagnostics

Fig 4.1: The schema for a semi-systematic review of microarray comparison papers

Fig 4.2: Defining sensitivity and specificity

Fig 4.3: Comparing performance of copy number microarrays

Fig 4.4: Sensitivity and specificity of copy-number detection varies between different microarrays

Fig 5.1: Reconstructing the clonal heterogeneity of cancer

Fig 5.2: Discovery of Mendelian disease causing *de novo* mutations using exome sequencing

Fig 5.3: Confirming site of tumour origin using ctDNA

Fig 5.4: Tracking tumour dynamics using ctDNA

Fig 6.1: Redefining breast cancer

Fig 6.2: Efficacy rates for different therapeutics

Fig 6.3: The potential for trastuzumab treatment

Fig 6.4: The potential for vemurafenib treatment

Fig 6.5: The potential for erlotinib treatment

List of tables

Table 4.1: mRNA comparison papers

Table 4.2: miRNA comparison papers

Table 5.1: Preparation of ctDNA exome libraries is possible with variable inputs of ctDNA

List of accompanying material

Publications submitted for this PhD by publication: This thesis is based on the following manuscripts, which are listed chronologically and referred to as e.g. Jennings *et al*¹ and/or their reference numbers within the text. Letters of support from co-authors are reproduced in Appendix 1 at the end of this thesis. The original publications are reproduced with the kind permission of the journals concerned, in Appendix 2.

Jennings et al. **A differential PCR assay for the detection of c-erbB 2 amplification used in a prospective study of breast cancer.** Mol. Pathol. 1997; 50(5):254-6¹.

Schmidt et al. **ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions.** Methods 2009; 48(3):240-8².

Curtis et al 2009. **The pitfalls of platform comparison: DNA copy number array technologies assessed.** BMC Genomics 2009; 10:588-611³

Git et al. **Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression.** RNA 2010; 16:991-1006⁴.

Lynch et al. **The cost of reducing starting RNA quantity for Illumina BeadArrays: A bead-level dilution experiment.** BMC Genomics 2010; 11:540-9⁵

Aldridge and Hadfield. **Introduction to miRNA Profiling Technologies and cross-platform comparison.** Methods Mol Biol 2012; 822:19-31⁶.

Curtis et al 2012. **The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups.** Nature 2012; 486:346-52⁷.

Forsheew et al. **Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA.** Sci. Transl. Med. 2012; 4(136): 136ra68⁸.

Murtaza et al. **Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA.** Nature 2013; 497:108-12⁹.

Idris et al: **The role of high-throughput technologies in clinical cancer genomics.** Expert Rev. Mol. Diagn. 2013; 13(2), 167-81¹⁰.

Azizan et al: **Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype**

of adrenal hypertension. Nature Genetics, 2013¹¹.

Hadfield & Eldridge: **Multi-genome alignment for quality control and contamination screening of next-generation sequencing data.** Frontiers in Genetics 2014; 20(5): 31¹².

Acknowledgments

This thesis describes work undertaken between 1996 and 2014, as such I have almost certainly missed out individuals who should have been acknowledged, for their direct input to this work, or for simply being fun to work with and/or supportive, and to those individuals I apologise. Thanks to my supervisors Mohammad Hajihosseini (School of Biological Sciences, UEA) and Barbara Jennings (Norwich Medical School, UEA). I must thank my wife Yvette for her constant support and encouragement in my career; also her, Holly's (10 ³/₄) and Sidney's (8) putting-up with my absence on many evenings whilst writing this thesis – I love you all very much. I'd not even have made it to University without the support of my extra-mum Anne without whom I know not where I would be today! I would not have started this PhD without the encouragement of Barbara my first boss, erstwhile supervisor, and continuing friend. I may not have allowed myself to be steered into starting the whole thing without the encouragement and support of John Wells, my boss at CRUK-CI. Some fantastic people, particularly at CRUK-CI, were behind the papers submitted as part of this thesis; including the groups of James Brenton, Carlos Caldas, Duncan Odom, and Nitzan Rosenfeld. The list of co-authors and colleagues is too long to include here but particular thanks goes to Matt, Rory, Mike, Tim, Murtaza, Feung, Andy, Inma & Sarah for being so great to work with. Thanks to my team at CRUK-CI who've been great to work with: Sarah (again), Nik, Michelle, Claire, Sara, Christian, Inma (again), Hannah, Fatimah, Sarah and Ros. Lastly thanks to Francis (RIP), James and Fred (RIP) without whom God only knows what any of us would have done?

When I started this thesis the people I spoke to seemed genuinely surprised that I had not completed a PhD already, thanks: hopefully you'll be reading this without major revision or delay!

“Of the three main activities involved in scientific research, thinking, talking, and doing, I much prefer the last and am probably best at it. I am all right at the thinking, but not much good at the talking”

Fred Sanger 1918-2013.

Chapter 1: Introduction

Genomics technology

The completion of the Human Genome Project (HGP, discussed in more detail below) provided the tools and knowledge to apply genome sciences to human disease. The completed sequence underpinned rapid developments in high-throughput analysis technologies such as microarrays (discussed in chapter 2), however sequencing, specifically next-generation sequencing, is the dominant tool in biological research today.

DNA sequencing

Sequencing as a technology was first applied to proteins when Fred Sanger described a new method for the identification and estimation of the free amino groups of proteins and peptides¹³. At about the same time Pehr Edman published the label-cleavage method for protein sequencing later termed “Edman” degradation¹⁴. The complete sequence of the Insulin B chain was published in two papers by Sanger and Tuppy in 1951 in the *Biochemical Journal*^{15,16}, for which Sanger received the 1958 Nobel prize for Chemistry. In 1968 Sanger published a method for RNA sequencing that used two dimensional fractionation of degraded RNA molecules on paper¹⁷. Large RNA molecules were enzymatically degraded, the degradation products were separated by chromatography and their sequences determined, the original RNA sequence was assembled from the multiple shorter sequences; in effect this was the first demonstration of shotgun sequencing. In 1975 Sanger and Coulson published the “plus and minus” method¹⁸ of DNA sequencing, which also used electrophoretic separation, and was a significant improvement on previous techniques. Sequencing was limited to 80bp making sequencing anything much larger laborious. However this was the method used to determine the first complete genome: that of bacteriophage PhiX¹⁹. In 1977 Sanger, Nicklen and Coulson published the “dideoxy” or “Sanger” method for sequencing that is still in use today, and for which Sanger received the 1980 Nobel prize for Chemistry²⁰. Sanger sequencing is an *in vitro* method that uses a DNA polymerase to copy a DNA template in the presence of chain-terminating dideoxynucleotide triphosphates (ddNTP) that are incorporated randomly into the growing DNA strand. These ddNTPs lack the 3'-hydroxyl group and cannot be extended by addition of another base, producing a final reaction containing molecules that differ in length by a single nucleotide, and that can be separated by gel-electrophoresis allowing the original DNA sequence to be read from the order of fragments in the gel. The original method required four reactions to be performed using a single radiolabelled dideoxynucleotide in each, the four reactions were run in four separate lanes on an acrylamide gel allowing DNA sequences of a few hundred bases to be read. Companies like Applied BioSystems significantly improved Sanger sequencing during their commercial development. Fluorescently labeled four-colour systems allowed reactions to be run in

parallel; dye-primer sequencing²¹ attached the fluorescent molecule to the sequencing primer and produced very even peak heights, but required four reactions to be performed independently and was limited to specific priming of reactions, dye-terminator sequencing²² attached the fluorescent molecule to the ddNTPs allowing a single sequencing reaction. The first DNA sequencers automated slab-gel electrophoresis, but a series of technical improvements led to the capillary electrophoresis sequencing instruments like the ABI 37030XL^{23,24}, the workhorse of DNA sequencing providers today.

Next-generation sequencing

The term next-generation sequencing (NGS) applies collectively to methods developed to replace Sanger sequencing. A 2004 review by Shendure *et al* stated the main reasons why more sequencing was required after completion of the HGP, and described the methods being developed²⁵. In 2005 Jonathan Rothberg and colleagues published the 454 system, demonstrating how in a single run they could sequence the genome of *Mycoplasma genitalium*²⁶; and in 2008 they published the 7-fold coverage genome of James Watson²⁷, which took just two months to complete at a cost of around \$100,000, discovering 3.3 million SNPs, 40,000 InDels and copy number variations. Due to the recent demise of Roche 454²⁶ (Roche, USA) there are just two main systems in use today Ion Torrent²⁸ (Life Technologies, USA) and Illumina²⁹ (Illumina Inc., USA); who have the lions share of users.

NGS methods fundamentally change the three major components of a sequencing experiment: library construction, template preparation and sequencing. Library construction uses an “*in vitro* cloning” workflow to replace plasmid cloning library preparation. Template preparation relies on emulsion PCR (Roche 454, Ion Torrent) or solid phase PCR (Illumina) to amplify single-molecules from the sequencing library, and replaces Big-Dye PCR²⁴. Sequencing of millions or billions of reads uses various methods: emulsion PCR and pyrosequencing²⁶ (Roche 454), emulsion PCR and semi-conductor sequencing²⁸ (Ion Torrent), or bridge-amplification and sequencing-by-synthesis²⁹ (Illumina, and described in more detail in chapter 2). Their development has been rapid, and the impact on biology has been dramatic with over 11,000 NGS publications since 2005 (PubMed: “next generation sequencing”). Many methods have been developed, some from earlier technologies, others that are unique to NGS and over 50 distinct methods have been reported³⁰.

NGS methods fall broadly into two types; those that rely on counting the number of reads that map to a specific genomic locus (e.g. methods to determine gene or exon expression levels, or DNA copy-number), and those that rely on determination of contiguous sequence (e.g.

methods to discover mRNA isoforms from full length assembled cDNA, or DNA structural variations such as translocations). Three NGS methods used in the publications submitted as part of this thesis are briefly described in chapter 2. The pace of development has not slowed, new methods continue to be published, some platforms are now obsolete (Helicos³¹ (Helicos, USA), SOLiD³² (Life Technologies, USA), and Roche 454²⁶ (Roche, USA)), and new platforms are being released Pacific Biosciences³³ (Pacific Biosciences, USA), Complete Genomics³⁴ (Complete Genomics, USA) and Oxford Nanopore (Oxford Nanopore Technologies, UK). There is also choice in how much and how fast, sequence data can be produced. The Illumina MiSeq, NextSeq, HiSeq and X-Ten (Illumina Inc., USA) systems all use the same library preparation and sequencing by synthesis, with slight variations in chemistry. They deliver very different amounts of sequence, gigabases to terabases, in different timescales, hours to days, and have very different operating costs. All of this allows users to select a platform that best fits their laboratory.

The Genome

A genome is the genetic material of an organism and is unique to that individual organism. In Humans the haploid genome is 3.2 Gbp (3.2 billion base pairs) in size. Proteins, primarily histones, which together with DNA are termed chromatin, organize the packaging of the genome into 22 autosomes and 2 sex chromosomes. The nucleosome, 140-150bp of DNA wound around a complex of eight histones, is the basic building block of chromatin. Chromatin also allows control of gene expression by allowing the “opening” and “closing” of DNA to make it accessible, or not, to polymerases and other proteins. Nucleosomes can be spread out like beads-on-a-string in “open” euchromatin containing actively transcribed genes; or further packaged into tightly-wound 30nm fibres as “closed” heterochromatin containing inactive and non-transcribed genes. The histones in chromatin can also be modified epigenetically to control gene expression, activating or repressing transcription. Trimethylation of histone 3 at lysine number 4 (H3K4Me3) is broadly correlated with high levels of transcription³⁵, whilst trimethylation of histone 3 at lysine number 27 (H3K27Me3) is a marker of gene repression³⁶.

Analysis of the human genome allows estimation of the number of genes, many of which do not make proteins

The human genome was completed in 2003 to an accuracy of 99.99% or one error in 10,000 basepairs³⁷. Only 50Mb (1.5%) is sequence that codes for proteins, 100Mb or more is regulatory, 50% is repetitive DNA, the term junk DNA has been largely eradicated. There are currently 19,942 protein coding genes^{38,39}, however large numbers of non-coding genes that

act directly through their transcribed RNAs, and psuedogenes increase the gene count to 58,688. The central dogma of molecular biology states that "DNA makes RNA and RNA makes protein" but this linear flow of information has proven to be over-simplified, and the relationships are not one-to-one but also one-to-many. As such the transcript count is currently 194,334 with 79,460 of those being protein-coding. All human cells share the same genome, however the genome is organized and regulated in such a way that allows it to create the distinct phenotypes we recognize as 1000 cell types. Up to 80% of human DNA is functional, i.e. has some form of biochemical role, from protein coding to simply transcribed⁴⁰. The genome is complex and there may be very little redundant DNA.

Genes are comprised of exons, introns and regulatory sequences

The 20,000 protein coding genes in the Human genome contain introns, exons and associated regulatory elements. Exons are on average 170bp in the Human genome^{41,42}. Each gene contains an average of just over 8 exons, however there is a wide variation in exon numbers: the largest Human gene *TTN* has 313 exons. The regulatory regions consist of DNA modifiers (including chromatin modification and DNA methylation), transcription modulators (including transcription factors, enhancers, activators, repressors and silencers) and translation modulators (including splicing and mRNA degradation). Differential splicing of genes, where exons are skipped resulting in variants of the protein, has been reported in up to 95% of multi-exon genes⁴³. Transcription is also complicated by the fact that it can occur in both sense (from the DNA strand the gene is annotated as occupying) and anti-sense (the opposite DNA strand), and that this happens across much of the non-coding genome^{44,45}.

The Human Genome Project

The sequencing of the complete human genome: the Human Genome Project (HGP), was first proposed in 1984 not quite ten years after the publication of Sanger sequencing²⁰. The publicly funded Human Genome Project formally began in 1990, a commercial project began in 1998 when Craig Venter started Celera's shotgun sequencing of the Human genome. The draft genome was completed in 2001 with joint publications in *Nature*⁴⁶ and *Science*⁴⁷, the completed genome (98% of the genome, at 99.9% accuracy) following in 2003. The formation of Celera Genomics in 1988 was seen as a challenge to the academic HGP, ultimately both were published at the same time, however the Celera genome⁴⁷ made significant use of the publicly available genome sequence data⁴⁸.

The HGP team had primarily used bacterial artificial chromosome genomic DNA libraries as the basis of sequencing. Clones were first physically mapped to the genome, allowing a

20

“tiling path” of clones to be selected for actual sequencing. Those selected were subcloned to produce smaller DNA fragments that could be sequenced. The shotgun method developed by Celera removed the physical mapping process making sequencing more efficient, but requiring more computational resources. The shotgun method has become the preferred method for sequencing genomes.

During the ten year project sequencing cost per base dropped over 100 fold. This was due to a large investment in technology development by the funders, sequencing teams and equipment providers. The Sanger sequencing method itself was significantly improved; as were cloning methods, sample preparation, electrophoresis, instrumentation and software, all of which contributed to the dominant method today: Applied Biosystems BigDye chemistry and 3730XL sequencer (Applied Biosystems, USA)²³. Ultimately this investment in technology would lead to the next-generation of DNA sequencing methods.

Cancer

Cancer is a disease of the genome and its underlying cause is germline or somatic mutation. DNA damage occurs at a rate of many thousands of single- and double-strand breaks per day⁴⁹. Most damage is fixed by the cells DNA repair processes but a small number of mutations are maintained, and these may confer a survival advantage on the mutated cell, driving a Darwinian evolutionary process that underlies the development of many cancers⁵⁰. This development may take many years. Bert Vogelstein first proposed that tumourigenesis occurs by the sequential mutations of oncogenes and tumour suppressor genes in 1990^{51–53} (**Fig 1.1**). Mutations required for tumourigenesis are termed “driver mutations”, cancer cannot develop without them, but uncontrolled cell division allows other mutations to accumulate that are by-products of tumourigenesis, these are termed “passenger mutations”. Eventually a critical mass of mutational load causes a tumour to develop⁵⁰. Several key biological pathways are commonly perturbed during tumourigenesis; these have collectively been termed “Hallmarks of Cancer”⁵⁴, and provide a conceptual framework for understanding the complexity of cancer biology (**Fig 1.2**). They highlight the fact that many targeted therapies ultimately fail due to biological redundancy in the hallmark pathways. A molecularly targeted therapy may not completely inhibit all tumour cells, allowing a few cells to adapt to the selective pressure imposed by treatment.

Cancer is a heterogeneous set of diseases

There are around 60 types of cancer, each with different causes, symptoms and treatments. Cancers can have distinct sub-types, which may have specific treatments and prognosis.

Ultimately cancer is a disease of an individual's genome and requires personalised treatment. There are around 300,000 cancer diagnoses in the UK each year, and around 1 in 3 people will develop cancer during their lifetime, but over one third of these will be in people over the age of 75, and only 2% of cancers will occur in people under the age of 24. Around half of new cancers occur in the breast, lung, prostate or bowel. Around 50% of cancer patients will live over ten years after diagnosis, and cancer survival rates have doubled in the last 40 years; but cancer still accounts for 1 in 4 deaths in the UK with 20% of those deaths being due to lung cancer. Lifestyle choices such as smoking, alcohol, diet and obesity are linked to 40% of an individual's lifetime cancer risk⁵⁵.

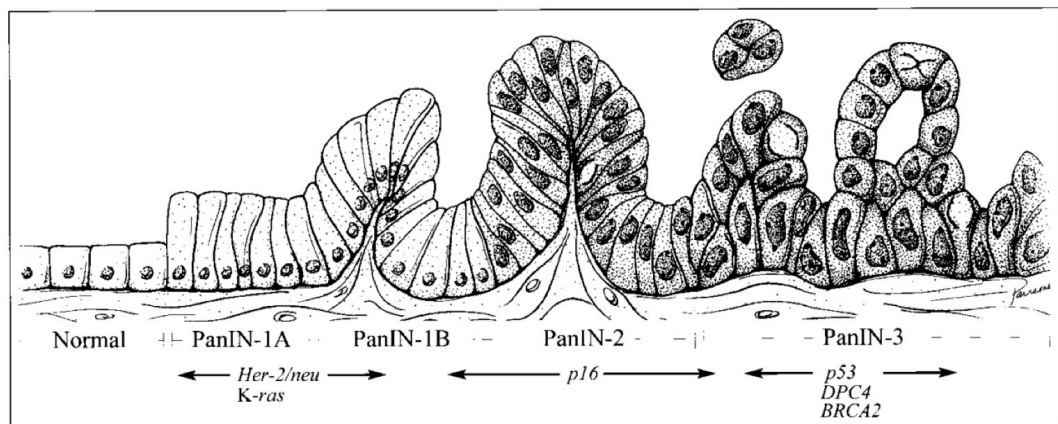
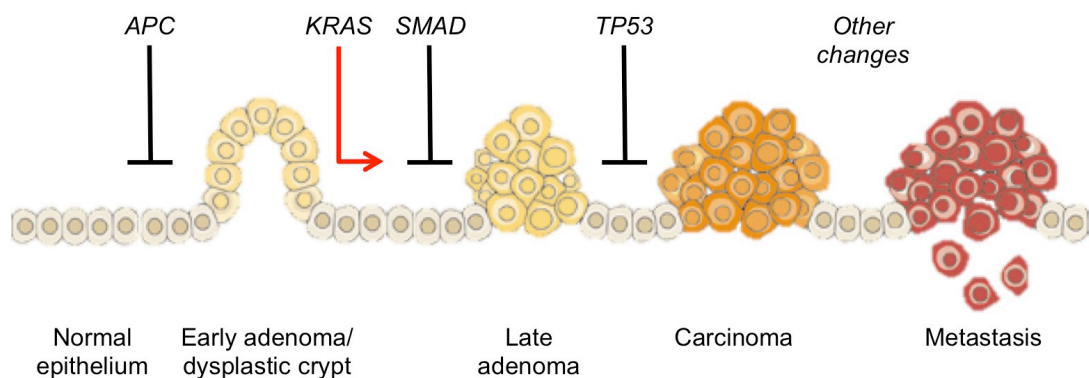
Breast cancer is the second most common cause of cancer mortality in women

Breast cancer is second only to lung cancer in female cancer mortality, accounting for over 11,000 deaths per year (15%). However our understanding of the causes of breast cancer is still developing. The heterogeneity of disease, with four main sub-types: Luminal A, Luminal B, HER2+ and triple-negative, makes finding the driver genes for breast cancer a challenging task. Additionally intra-tumour heterogeneity has been revealed as important in determining outcome, and the molecular stratification of breast cancer patients is likely to lead to better patient outcomes and improved biological analyses. Although there are several germline genetic risk factors for breast cancer, the majority of lifetime risk is due to somatic mutation. The two major susceptibility genes, BRCA1 and BRCA2 increase risk by up to 80% but are rare in the general population. The outcome for breast cancer patients is dependent on the several factors: type, stage, grade.

The cancer Genome

The relatively recent publication of the first cancer genomes^{56,57}, and the early results from the International Cancer Genome Project⁵⁸⁻⁶² (ICGC) demonstrate how much can be learned from sequencing cancer genomes. The ICGC aims to generate a comprehensive catalogue of somatic mutations in 50 cancers and in 500 cases from each, generating 25,000 cancer genomes. In the UK data from 200 esophageal cancers has recently been published⁶³; and 500 ER+, *ERBB2* amplified breast cancers are being analysed for all classes of mutations, genome-wide DNA methylation, and RNA expression. The UK is also delivering the bone, chronic myeloid disorders and prostate cancer ICGC projects. The ICGC brings together the Wellcome Trust Sanger Institute's: Cancer Genome Project, and the USA National Institute of Health's: The Cancer Genome Atlas projects. The merger of two large projects aims to maximise the dissemination of data and reduce duplication of efforts.

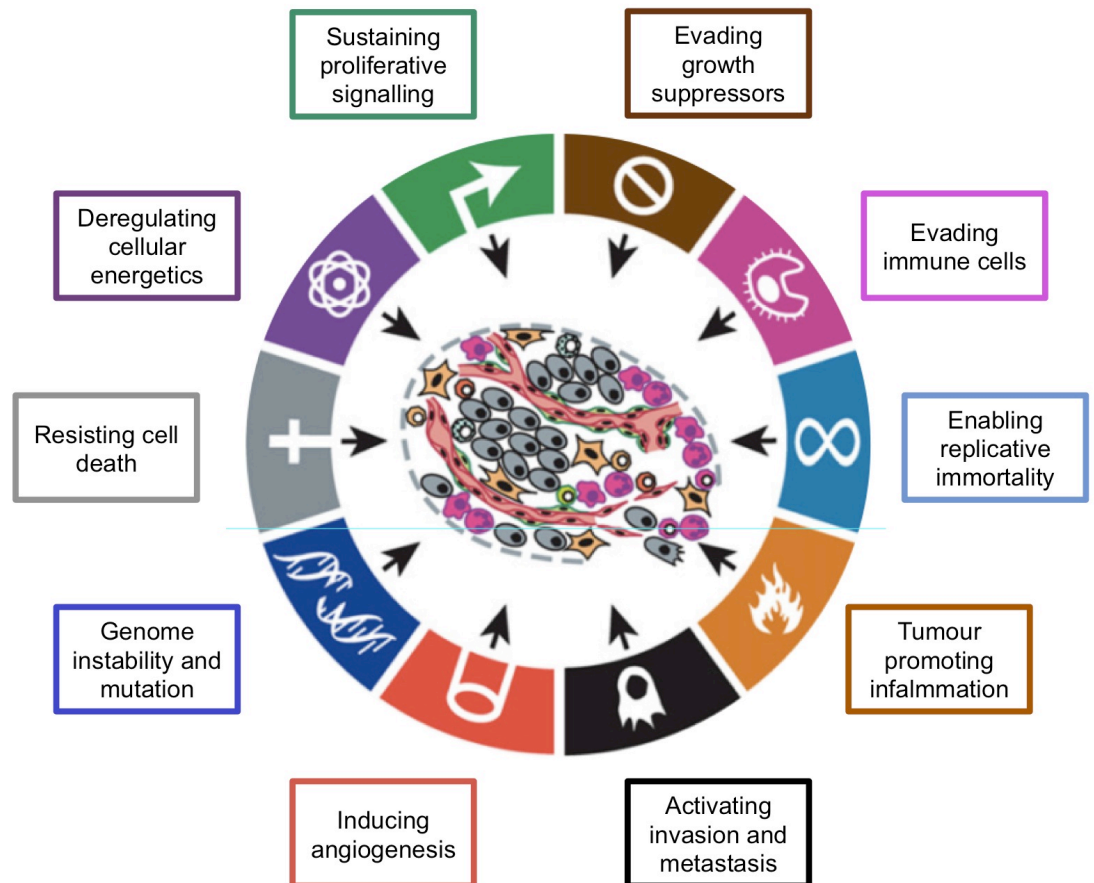
Fig 1.1: Cancer is caused by sequential mutations of specific oncogenes and tumor suppressor genes



Both colorectal and pancreatic cancer demonstrate the “Vogelstein model” of tumour progression from normal tissue, through various stages to cancer and finally metastases. Both loss of tumour suppressor genes and activation of oncogenes occurs. The accumulation of mutations happens over time.

Adapted from Knudson **Nature Rev Cancer** 2001 ⁽⁵²⁾ and reproduced from Wilentz **Cancer Res** 2000 ⁽⁵³⁾

Fig 1.2: The hallmarks of cancer simplify the complexity of cancer biology by describing cancers driving pathways



The hallmarks of cancer are ten biological phenotypes acquired during tumour development, that provide a conceptual framework for understanding the complexity of cancer biology.

Reproduced from Hanahan & Weinberg *Cell* 2011 ⁽⁵⁴⁾

The cells in most solid tumours are not an isogenic population; in most cases there are multiple sub-clonal populations⁶⁴. These evolve from the very first tumourigenic cell(s) and share some of the early driver and passenger mutations, but they may also be quite distinct having many other private mutations⁶⁰. This intra-tumour heterogeneity was first studied using whole exome next-generation sequencing of primary renal cancers and their associated metastases⁶⁴, demonstrating that over 30% of mutations were not shared across all samples from an individual. This work also revealed a convergent evolution by distinct mutation of multiple tumour suppressor genes in separate samples from an individual. Data emerging from the ICGC is revealing how common some gene mutations are, not just in a specific cancer, but across multiple cancers.

Different mutational mechanisms underlie the causes of cancer

Tumorigenesis and tumour progression are driven by mutation of the normal genome. Most of the early driving events, as proposed in the Vogelstein model, are mutations in oncogenes (genes where mutation leads to gain of function) or tumour suppressor genes (genes where mutation leads to loss of function). A landmark publication by Sjöblom *et al*⁶⁵, analysed over 3 million PCR reactions across 13,023 consensus coding sequences to show that individual tumors accumulate an average of 90 mutant genes, that a small subset contribute to tumorigenesis and that 189 genes were mutated at high frequency across cancers. Mutation occurs in many forms and cancer can be driven by one or another, or a mix, these include: single nucleotide substitutions; single-base to large insertions or deletions (InDels); genomic rearrangements (translocations), copy number aberrations (amplification and deletion); and loss-of-heterozygosity. The type of mutational process and the order that mutations are gained in a tumour give rise to a mutational signature unique to each tumour clone⁶⁶.

Tumour suppressor genes and oncogenes drive cancer

Tumour suppressor genes are those where mutation leads to loss of function. Most tumour suppressors have roles repressing the cell cycle, in protecting cells against tumorigenesis, e.g. DNA repair, or in promoting apoptosis. Many familial risk factors occur in tumour suppressor genes including the BRCA1 and BRCA2 genes involved in DNA repair. Both alleles need to be lost to be tumorigenic and this generally occurs through deletion, loss-of-heterozygosity and *de novo* mutation of the second allele. This requirement for two hits in a tumour suppressor gene was first proposed in retinoblastoma⁶⁷. Other important tumour suppressors include TP53 and PTEN, which both play roles in regulating the cell cycle.

Oncogenes are those where mutation leads to gain of function. Mutations can lead to oncogene activation through many mechanisms; e.g. increasing gene copy number or

transcriptional activity. The major oncogenes in breast cancer are *ERBB2*, *EGFR*, and *CCND1*. Oncogenes have been successfully used as targets for drug development, resulting in trastuzumab (*ERBB2*), imatinib (*BCR-ABL*) and erlotinib (*EGFR*), amongst others.

Summary

The accumulation of data about the cancer genome using next generation sequencing technologies in projects such as the International Cancer Genome Consortium (ICGC) is radically changing our understanding of cancer. It is already beginning providing evidence that demonstrates how treatments for one cancer can be beneficial for small numbers of patients with very different diseases, e.g. Herceptin treatment of *ERBB2* amplified pancreatic cancer⁶². It took 30 years to develop the Sanger sequencing method published in 1977 from the earlier protein and RNA methods. The next 30 years saw the completion of the Human Genome Project and the introduction of next-generation sequencing technologies. What the next 30 years will hold is not certain but, with the ability to sequence whole cancer genomes we are very likely at the start of a revolution in medicine driven by the genome.

Chapter 2: Methods presented, Experimental design and Quality control

Introduction

The papers submitted as part of this PhD by Publication thesis describe the use of several genomic technologies. This chapter provides a brief introduction and historical overview of the methods cited. The detailed description of the methods used can be found in the presented publications, attached as appendix 1.

Experimental design is a critical step before starting laboratory experiments, and quality control is recommended for many steps during the collection of experimental samples and data, including the statistical analysis. The principles of experimental design were established in the 1930s by R. A. Fisher⁶⁸, and although they were not developed for high-throughput genomics experiments they are eminently applicable. Almost every lab-scientist has at some point been confronted with Fisher's "post-mortem" quote by a statistician, probably for good reason.

R.A.Fisher (1938) *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*

Quality control (QC) is a fundamental step that can be applied at almost any point in an experiment. In genomic experiments it is often vital that QC be carried out before biological analysis and interpretation. In this chapter I will focus on the experimental design of microarray experiments and the QC of next-generation sequencing experiments.

Methods Presented

Immunohistochemistry

Immunohistochemistry (IHC) is a method for detecting proteins in tissue sections, typically used in a pathology laboratory for clinical diagnostics, e.g. *ERBB2* via the HercepTest⁶⁹ (**Fig 2.1**). Tissue sections are incubated with antibodies directed against the protein of interest allowing the tissue and intra-cellular distribution and localisation of specific cellular components to be determined. IHC can be interpreted qualitatively but can also be quantitative or semi-quantitative.

Two limitations of IHC are considered within this thesis: first, the interpretation of results, which can be subjective affecting the validity and reliability of clinical tests⁷⁰; and second, formalin fixation, which can damage nucleic acids⁷¹. Samples for IHC are usually formalin-

fixed and paraffin embedded (FFPE) to stabilise tissue architecture. FFPE causes at least four different DNA modifications⁷² and two of these affect the nucleic acid analysis techniques described in this thesis: strand breaks and depurination caused by formaldehyde hydrolysis. The latter can be “rescued” to some extent and the use of uracil-DNA-glycosylase to reduce C>T transitions by hydrolysis of the uracil generated from cytosine deamination has been shown to have good effect⁷¹. Even so, the fixation process and the time since fixation have a dramatic effect on the usability of FFPE DNA in genomic analyses. Whilst many clinical studies do collect fresh frozen samples, the majority of clinical samples are processed by pathology laboratories and are formalin fixed.

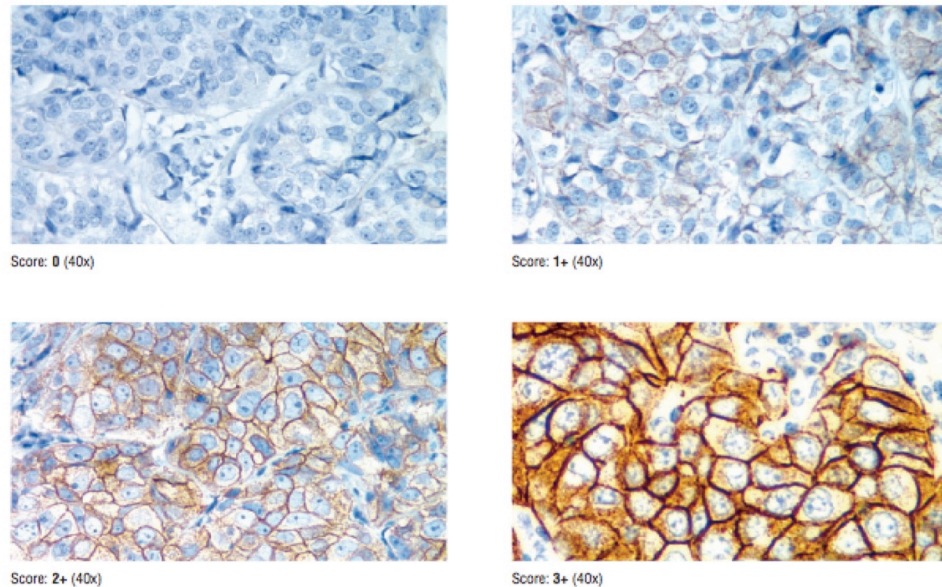
Real-time quantitative PCR

Quantitative real-time PCR (qPCR) was developed in the early 1990’s by teams at Roche, Genentech and Applied Biosystems^{73–75}. An adaption of standard PCR⁷⁶, qPCR allows simultaneous amplification and detection of specific DNA sequences. PCR consists of three phases: exponential, linear and plateau. During the initial PCR cycles, the signal generated by a fluorescent reporter cannot be distinguished from the background. However the signal begins to increase exponentially and then linearly before entering the plateau phase of PCR, where it stabilizes. The fluorescence data are plotted to generate an amplification curve, which is used to determine quantitative information. The point on the curve used to determine quantity values is referred to as the quantification cycle (C_q)⁷⁷ (**Fig 2.2A**). Its placement is somewhat arbitrary but should be consistent across an experiment. Gene expression or copy-number are determined by comparing the differences in the number of cycles where C_q is reached for test and control assays (ΔC_q) between test and control samples ($\Delta\Delta C_q$) (**Fig 2.2B**), or by comparison to a standard curve using a reference standard (**Fig 2.2C**). Assays must have similar PCR efficiencies to generate robust qPCR data⁷⁸.

Methods to detect amplicons generated by qPCR

Two methods are commonly used to detect PCR amplicons: non-specific intercalating fluorescent dyes e.g. SYBR® Green, or the use of fluorescently labeled oligonucleotides e.g. TaqMan probes (**Fig 2.3**). SYBR® Green (Molecular Probes, USA) intercalates with double-stranded DNA in a non-sequence specific manner and shows greatly enhanced fluorescence upon intercalation^{79,80}. SYBR® Green fluorescence is directly proportional to the amount of double-stranded DNA in the reaction.

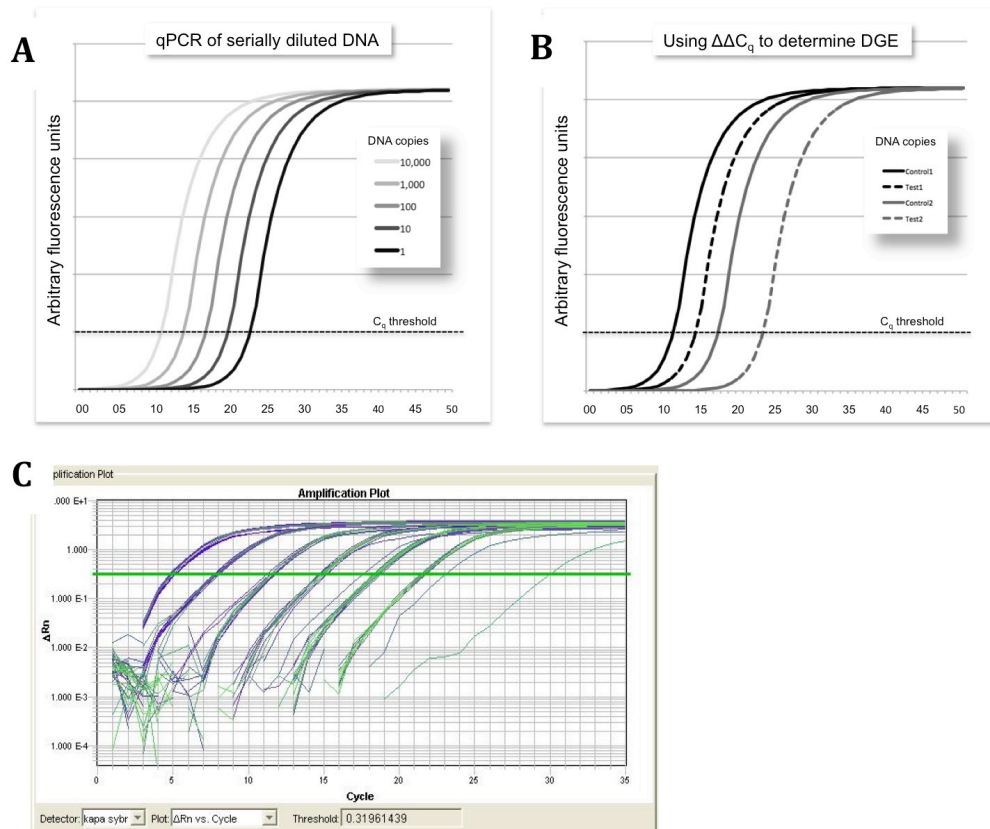
Fig 2.1: *ERBB2* amplification testing by IHC



Example HercepTest IHC results: After antibody incubation and chromogenic staining different levels of *ERBB2* protein are detectable. Samples showing no detectable staining are scored 0, Those with <10% of tumour cells, or with incomplete membrane staining are scored 1+. Those with >10% of tumour cells and with weak to moderate membrane staining are scored 2+. Those with >10% of tumour cells and with complete membrane staining are scored 3+.

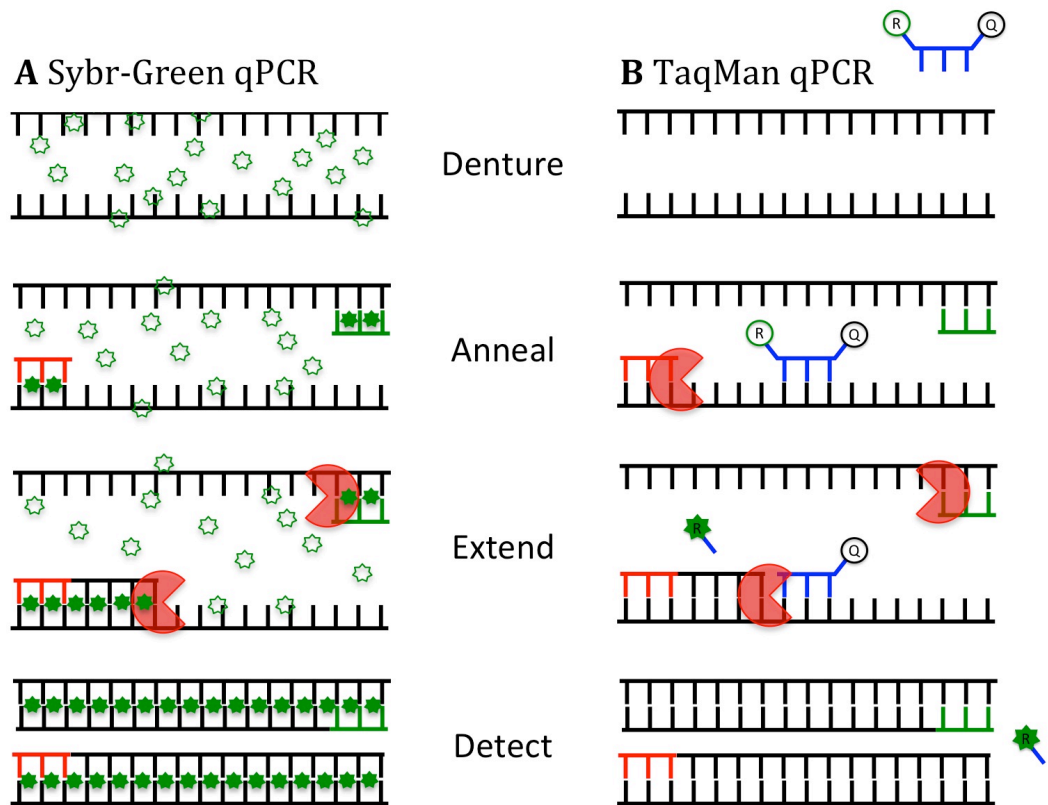
Reproduced from **Dako HercepTest Interpretation Manual – Breast 2002**⁽⁶⁹⁾.

Fig2.2: qPCR is a sensitive method for measuring nucleic acids



A: A ten-fold serial dilution of input DNA. The C_q threshold is set just above the exponential phase, $3C_q$ separate each ten-fold diluted sample, e.g. a two-fold change in expression results in a $4C_q$ difference. **B:** The $\Delta\Delta C_q$ method measures the difference of differences in C_q values of control and test genes in two conditions: $\Delta\Delta C_q$ ($6-2=4$) = Test1 ΔC_q ($14-12=2$), Test2 ΔC_q ($25-19=6$). **C:** An example of a standard curve from a quantification of next-generation sequencing libraries using six replicates for each data point, note the good overlap of replicates.

Fig 2.3: Two different qPCR detection methods



Both SYBR green and TaqMan probe qPCRs go through the same PCR cycling. **A.** SYBR-green does not fluoresce in the presence of denatured DNA, during annealing and extension the SYBR dye intercalates and fluoresces but detection is not performed until extension is complete. **B.** TaqMan probes cannot fluoresce due to the presence of the quencher molecule, the fluorescent reporter quencher is cleaved during extension from its quencher by the 5'-3'-exonuclease of Taq polymerase and fluoresces, detection can be performed during or after extension.

TaqMan hydrolysis probes^{81,82} bind to specific DNA sequences and contain fluorescent reporter and quencher molecules in close physical proximity, inhibiting reporter fluorescence. During PCR extension, the 5' exonuclease activity of Taq polymerase cleaves the probe removing proximity inhibition. The fluorescence from the reporter is directly proportional to the number of PCR products amplified and probe molecules cleaved. Both methods work well for relative quantification of gene copy-number or mRNA gene expression. SYBR® Green is an easy, cost-effective and sensitive method to implement. It can also be used for qualitative, and mutation detection analysis using a “melt-curve”. However, SYBR® Green detects all double-stranded DNA in a PCR and can be affected by “primer-dimer” and other PCR artefacts, and SYBR® Green assays cannot be multiplexed. TaqMan assays consist of three oligonucleotides, a pair of standard PCR primers and a fluorescently labeled TaqMan probe, as such TaqMan assays are significantly more expensive than SYBR® Green. The addition of a sequence specific probe can significantly increase specificity and detection of non-specific amplicons is less problematic. TaqMan probes can also be multiplexed.

Standardising qPCR experiments

The success of real-time PCR has led to thousands of publications, however many of these do not report detailed methods and a few use methods that are sub-optimal. To address this the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines were published⁷⁷. These aim to encourage users of qPCR to include detailed information about sample collection and extraction, RNA quality, reverse-transcription methodology, PCR efficiencies, and analysis parameters.

qPCR as a diagnostic tool

One of the earliest qPCR papers to report the use of TaqMan probes, used *ERBB2* gene amplification as an example of quantitative gene analysis⁸³. The first commercial qPCR-based *ERBB2* test was released in 2001 by Roche⁸⁴, which improved on previous methods by multiplexing *ERBB2* qPCR with a reference gene co-localized on chromosome 17 to take into account possible Chr17 polysomy. The test has not been reported in peer-reviewed literature. Other *ERBB2* tests have been published⁸⁵, and one comparison of qPCR to IHC demonstrated high correlations⁸⁶. qPCR plays an increasingly important role in diagnostic testing because it is sensitive and objective as well as rapid and cost-effective. Its use in diagnostics was reviewed for *ERBB2* and *TOP2A* amplification analysis, and Epstein–Barr and human papilloma virus involvement, in cancer diagnostics⁸⁷.

Microarray

The analysis of DNA, and later RNA and Protein using Southern⁸⁸, Northern⁸⁹ and Western⁹⁰ blotting respectively allowed only single-target analysis. These approaches were based on hybridising radioactively labeled oligonucleotide probes to a membrane bound sample, or arrays of samples. Schena *et al* described the adaptation of this approach achieved by simply reversing the system and hybridising fluorescently labeled samples to cDNA probes immobilized on glass microarrays⁹¹ (**Fig 2.4**). At the same time several other methods were introduced: Affymetrix used photolithographic *in-situ* synthesis of oligonucleotides in place of PCR-amplified cDNA clones^{92,93}, Agilent used ink-jet deposition for *in situ* oligonucleotide synthesis⁹⁴, and Illumina attached oligonucleotides to microscopic beads deposited onto silicon slides to create randomly ordered microarrays^{95,96}. These approaches are now used for whole-genome analysis of copy-number variation (CNV) and loss-of-heterozygosity (LOH) or whole-transcriptome analysis of differential gene expression (DGE).

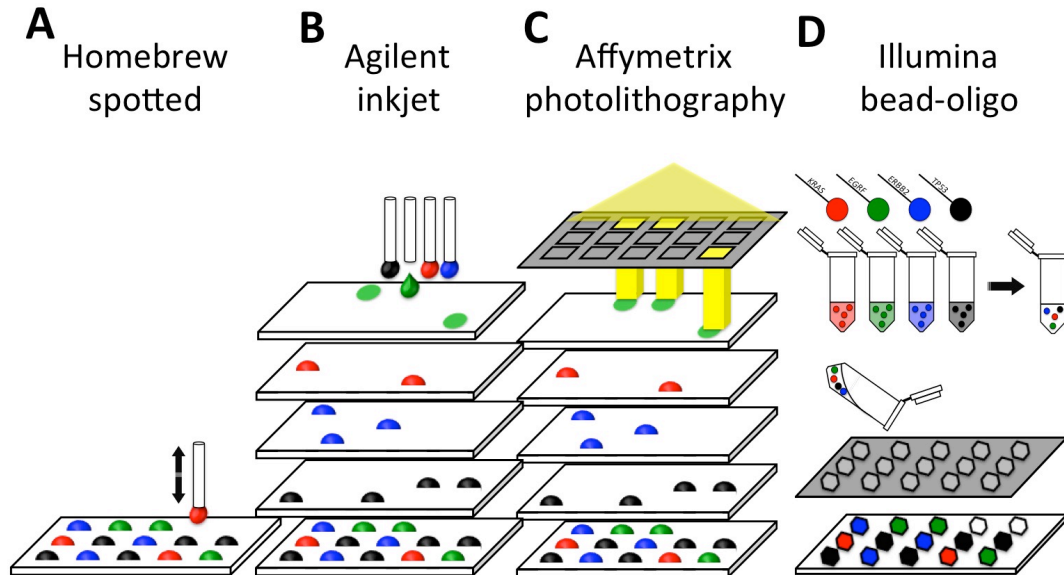
There are three major issues related to the use of microarrays. First, the use of probes restricts analysis to regions of the genome of known or inferred sequence; second, cross-hybridisation between homologous sequences confounds data analysis; third, design of probes can be complicated by both of these factors. However the quality of microarray data today is exceptional, the data analysis tools are mature and easy to use and many researchers still benefit from using microarray over maturing next-generation sequencing technologies.

Microarray analysis of RNA: differential gene expression (DGE) and micro-RNA

DGE analysis of mRNAs generally involves oligo-dT enrichment of mRNAs, which are converted to cDNA, or secondarily converted to cRNA whilst incorporating fluorescent- or biotin-labeled nucleotides. Labeled samples are then applied to whole-genome microarrays as single- or dual-colour hybridisations and processed for laser scanning. Finally, signal intensities are converted to absolute and/or differential gene expression measurements (**Fig 2.5**).

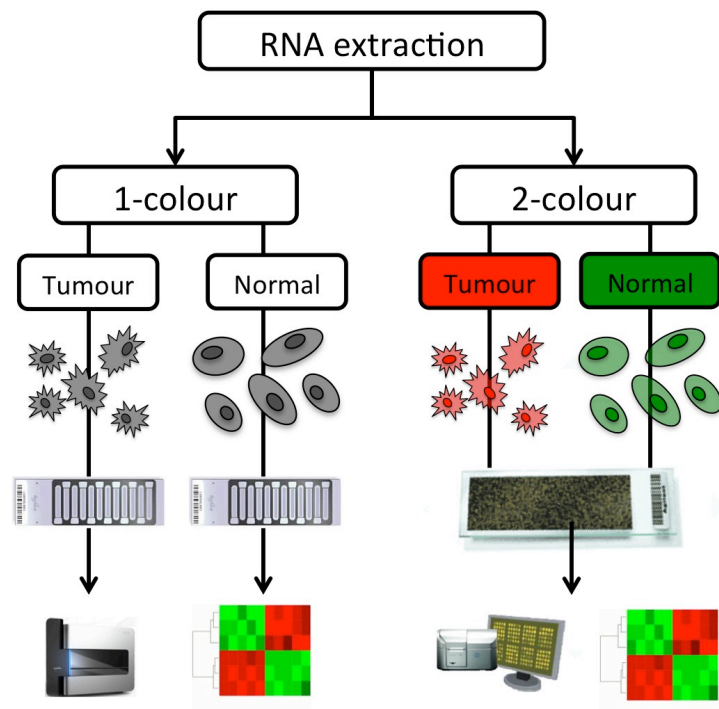
MicroRNAs (miRNAs) are 20-23 base pair long non-coding RNA molecules. They regulate gene expression by binding to target mRNAs and repressing protein translation or directing mRNAs for degradation. miRNAs are generated by cleavage of longer transcripts: primary-miRNA to precursor-miRNA to miRNA. The miRNA database miRbase⁹⁷ contains over 24,000 miRNAs, including 2000 annotated with high-confidence⁹⁸. Micro-RNA analysis by microarray is similar to mRNA analysis but complicated by their very short length, high

Fig 2.4: Methods for microarray manufacture



The four major methods for microarray manufacture are indicated. **A.** Home-brew and commercial spotting onto glass slides is rarely used today. **B.** Agilent ink-jet printing *in-situ* synthesises oligonucleotide probes. **C.** Affymetrix photo-lithographically *in-situ* synthesises oligonucleotide probes. **D.** Illumina attaches oligonucleotide probes to beads that are randomly deposited into silicon arrays. The three major commercial methods share similarities in manufacture and processing, all are oligonucleotide based. All methods produce microarrays with similar, or identical probe characteristics, as indicated by the bottom-most cartoon.

Fig 2.5: Microarrays for mRNA differential gene expression



One- and two-colour mRNA microarray analysis results in the same differential gene expression results (as indicated by heatmaps). In one-colour systems RNAs are labeled with the same dye and hybridised to separate arrays, the absolute intensity of probes is used to generate differential gene expression (DGE) results. In two-colour systems RNAs are labeled with the red and green dyes and hybridised to the same array, the ratio of probe intensities is used to generate DGE results.

homology, sequence specific biases in RNA ligases, and the need to discriminate pri-, pre-, and mature-miRNAs. Designing hybridisation probes to accurately discriminate miRNAs is a difficult task particularly due to the variable T_m of such short molecules (reviewed in^{6,99}).

Microarray analysis of copy-number variation (CNV)

Microarray CNV analysis is achieved by labeling and hybridising genomic DNA to microarrays. It was developed from comparative genomic hybridization (CGH) techniques¹⁰⁰ and originally called array-CGH (aCGH). In a typical CNV experiment, tumour and normal samples will be directly compared. An important advance in CNV, and vital for LOH analysis was the development of SNP genotyping arrays from Affymetrix and Illumina. Copy number on these platforms is inferred from probe signal intensities, similar to more conventional aCGH and as such is referred to as snpCGH. The use of SNPs allows genotype calls to be made and as such allows LOH to be analysed by identifying regions of the genome where only one parental copy is present^{101–103}. As cancer can be driven by LOH as well as CNV, deciphering both has significant advantages as demonstrated by Curtis *et al* 2012⁷ and discussed in chapter 5. The methods for analysis of DNA are more complex and varied than those for RNA analysis and can include degenerate PCR or whole-genome amplification reactions that incorporate fluorescent- or biotin-labeled nucleotides, or single-primer extension from millions of oligonucleotide probes. Microarrays are processed for laser-scanning and signal intensities are converted to absolute and differential CNV measurements (Fig 2.6).

Microarray as a diagnostic tool

A landmark paper by Golub *et al* in 1999 described the classification of acute myeloid and lymphoblastic leukaemias based on gene-expression signatures¹⁰⁴, demonstrating the utility of microarrays in cancer diagnostics by automatically determining the class of new leukaemia cases. Many other molecular classifiers have been published but few have been implemented in the clinic. Of those that have the PAM50 molecular classification of breast cancer initially reported by Perou *et al*¹⁰⁵; the Oncotype DX test, a 21 gene molecular classifier for quantifying the risk of disease recurrence; and the MammaPrint test, a diagnostic test to assess the risk that a breast tumor will metastasize to other parts of the body approved by the FDA in 2007^{106,107}; are the most commonly reported. All aim to stratify patients into sub-groups for prediction or prognosis. The current status of the clinical applications of aCGH & NGS was reviewed by the author in Idris *et al* 2013 and submitted as part of this PhD by Publication¹⁰.

Next-generation sequencing

The work presented in this thesis has made extensive use of the Illumina sequencing technology, the success of which has been primarily down to the quality and quantity of sequence data, and the ease of use of the library preparation technology. This can also be seen in the large number of NGS publications (PubMed: “next generation sequencing” returns over 11,000 publications), as well as specific methods publications, e.g. RNA sequencing (over 3,000 PubMed articles), since the widespread adoption of NGS in 2010. And the potential for impact in the clinic is apparent by the rapid take-up of exome sequencing (over 3,000 PubMed articles), with 90% of papers published since 2012. Other NGS technologies and methods are not described here, but have been extensively reviewed elsewhere^{25,108,109}.

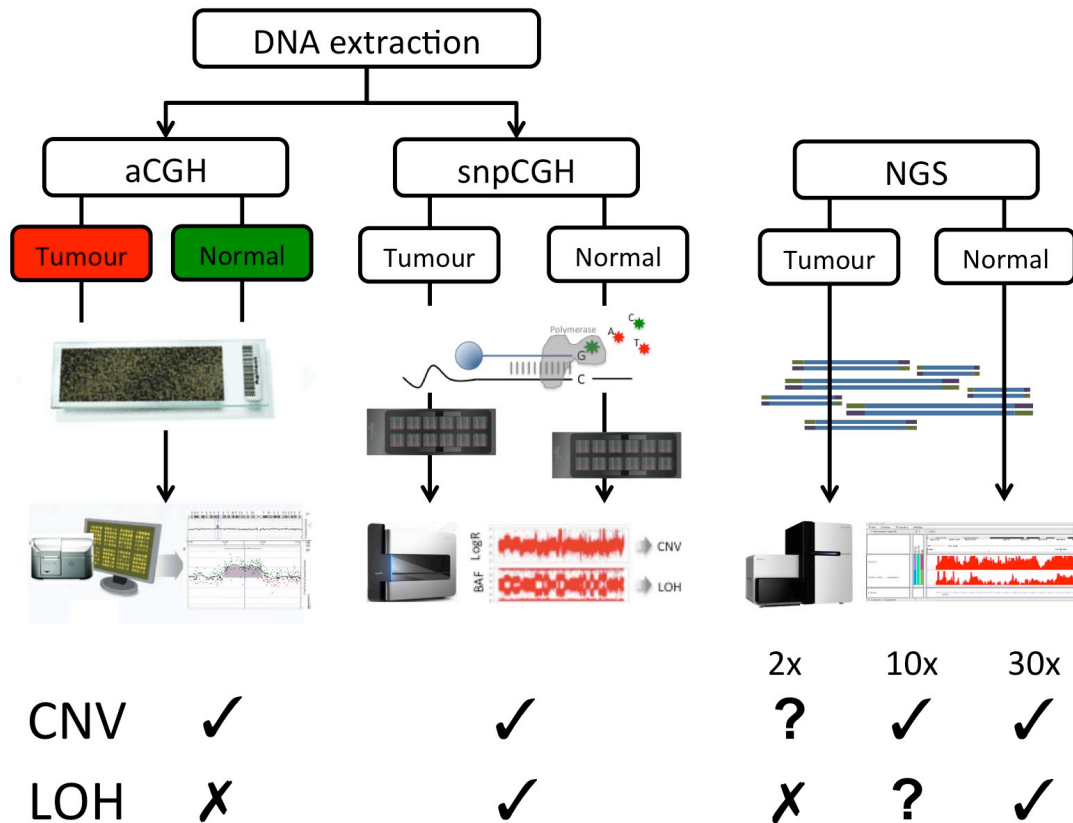
Illumina sequencing by synthesis

The Illumina sequencing-by-synthesis technology²⁹ is the most widely adopted NGS system (<http://www.omicsmaps.com> lists nearly 70% of NGS systems as from Illumina). Whole genome library construction uses fragmented DNA, to which Y-shaped adapter molecules are ligated, resulting in DNA strands with different adapters at each end. DNA libraries are denatured and hybridised to “flowcells” - glass slides coated with oligonucleotides complementary to the sequencing library adapters. Library molecules are amplified in a solid-phase PCR termed bridge-amplification, to form clusters of around 1000 molecules. Each cluster will generate a sequence read. Sequencing-by-synthesis uses reversibly-terminated fluorescently-labeled nucleotides. Each sequencing cycle consists of 3 steps: incorporation, imaging & cleavage. First, nucleotides are incorporated into the growing DNA strand, then each cluster is imaged to determine which base has been incorporated before the fluorescent label and blocking groups are chemically cleaved leaving the nucleotide ready for the next sequencing cycle (**Fig 2.7**). The current v4 SBS chemistry on HiSeq 2500 produces around 4 billion sequences of 125bp from each end of the DNA molecule.

RNA-sequencing (RNA-seq)

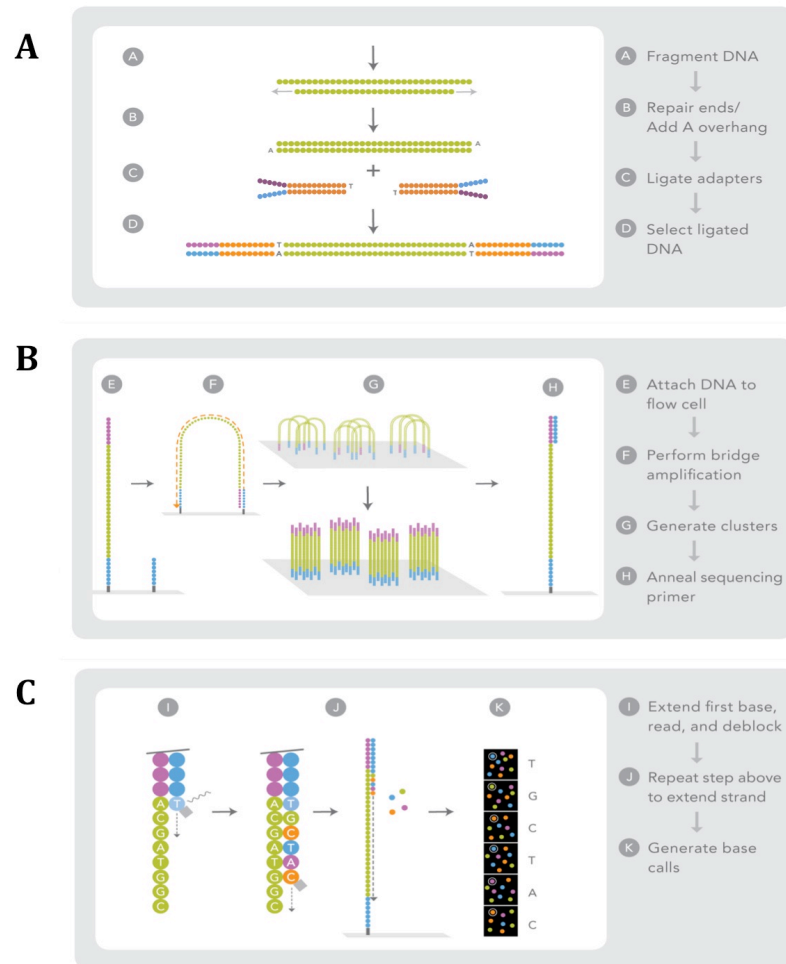
RNA-seq¹¹⁰ can be used to detect and measure coding and non-coding RNAs, differential gene expression, allele specific expression, splicing analysis and RNA editing studies. RNA-seq methods are varied and have been extensively reviewed and compared^{111,112}. There are two main methods: cDNA adapter ligation (the predominant method for mRNA analysis) or, direct RNA-ligation of adapters (used primarily for non-coding studies). The predominant method involves mRNA enrichment with oligo-dT, mRNA fragmentation, random-primed cDNA synthesis incorporating dUTP into the second-strand, adapter ligation, dUTP

Fig 2.6: Microarrays or NGS for CNV & LOH



One- and two-colour CNV microarray analysis results in the same differential copy-number results. Two-colour aCGH uses DNAs labeled with the red and green dyes and hybridised to the same array, the ratio of probe intensities is used to generate copy-number calls. One-colour snpCGH uses comparison of genotyping array intensities to infer copy-number, and also determine loss-of-heterozygosity. NGS data can be used to determine CNV and LOH but the depth of sequencing limits the analyses that can be performed.

Fig 2.7: Illumina next generation sequencing technology



Illumina sequencing is comprised of three steps. **A)** oligonucleotide adapters are ligated to fragmented DNA, enabling bridge-PCR and sequencing in the later steps. **B)** Single molecules are hybridised to a solid-surface and amplified by bridge-PCR to form colonies for sequencing. **C)** Sequencing-by-synthesis is a cyclical process where fluorescently-labelled and blocked nucleotides are incorporated, and detected; the block and label are then removed ready for the next cycle.

degradation and PCR. This results in strand-specific mRNA-seq libraries that can be used for differential gene expression. Sequencing of 10-20 million single-end 50bp reads is sufficient to generate data of the same quality as a microarray¹¹³. Sequencing reads are aligned and counted to determine differential gene expression. RNA sequencing is rapidly replacing gene expression microarray as the standard method for RNA analysis for several reasons: (i) RNA-seq provides a digital read-out of counts of RNA molecules; (ii) Read-depth can be adjusted to give the dynamic range required; and (iii) there is no need to limit analysis to pre-designed probes, and there are no cross-hybridisation artifacts to consider.

Chromatin Immunoprecipitation sequencing (ChIP-seq)

ChIP-seq^{114,115} is used to detect and measure DNA:Protein interactions. ChIP-seq methods have remained relatively unchanged since the first application of ChIP to microarrays¹¹⁶. DNA is cross-linked with its bound proteins by formaldehyde treatment, cells are homogenized, and chromatin is sheared before immunoprecipitation with an antibody to the protein of interest. This ChIP DNA is used to produce an NGS library for sequencing. Short reads are aligned to the genome to determine where the protein was bound. This method produces the characteristic peaks of ChIP-seq.

Exome-sequencing (Exome-seq)

The exome is the coding portion of the genome and has become a powerful tool for clinical research and now treatment¹¹⁷, as it requires comparatively little sequencing to a whole genome, at about 10%-30% of the cost. Whole genome libraries are prepared as described above then used in a capture hybridisation with single-stranded biotinylated oligonucleotide probes, designed to recognise each exon. These probes pull-out the exonic fragments only from the whole genome library and are then ready for exome sequencing.

Experimental design

The main issues affecting the design of genomics experiments are replication and randomisation. Designing an experiment can be as simple as deciding on replication levels and discussing the merits of different methods that might be used where several are available (discussed in more detail in Chapter 4). More complex designs can introduce randomisation and blocking into an experiment to help control for confounding factors or batch effects. Blocking is the arranging of samples into groups that can be analysed together and is often used to control for dye-labelling in microarrays and sequencing lane within an NGS run. Randomisation involves the allocation of samples (patient in the context of clinical trials)

across treatment groups. However, introduction of randomisation into a ChIP-seq or RNA-seq experiment is less clear, but the technological complexity of the experiments makes randomisation something that should be carefully considered during sample collection, extraction and processing.

The impact of replication on experimental design

Replication is the repetition of an experimental condition so that its variability can be estimated; it is a key component of a well-designed experiment. However, one of the first questions asked by researchers in experimental design sessions is “how many replicates do I need.” Biologists generally discriminate between biological (good) and technical (not-so-good) replication. Biological replicates for breast cancer would be samples from different individuals. However some studies may consider a secondary tumour from the same individual, or even a separate biopsy from the same tumour, as biological replication. A technical replicate in this instance would be the primary sample run twice. For tissue culture the issue is slightly easier as it is impossible to have true biological replicates (the cell line was derived once only) and replications of the experiment (different passage, different day, etc) are considered biological. Technical replicates are not a substitute for biological replicates and should only be used if biological replication is not possible¹¹⁸. The only exception to this rule is when performing technology comparison studies, as the aim of the experiment is to understand technical bias or comparability. As such the use of technical replicates gives the most power to these studies. In a micro RNA⁴ comparison study (discussed in chapter 4) we only used technical replication and four replicates for each technology platform. In a copy-number comparison study³ (also discussed in chapter 4) we used technical and biological replication and two to three replicates for each technology platform.

One of the first groups to address the impact of replication on microarray studies¹¹⁹ made sample size calculations and estimated power using published datasets. They clearly demonstrated that experimental sample size depends on variance of gene expression within the samples being studied; the desired detectable fold change between sample groups; the power to detect this change (probability of not committing a Type II error); and a chosen false positive (Type I error) rate. Users are often unable to specify any of these without *a priori* knowledge and the most useful method to generate this is a pilot experiment (examples of which are described in more detail in Chapter 4^{3,4}).

Experimental factors affecting replication and experimental design

The number of replicates required is a function of the samples available. High-quality and high-input nucleic acids result in a better signal to noise ratio than low-quality and low-input samples. We investigated the impact of reducing RNA input into differential gene expression studies using microarrays⁵. In this study we were able to demonstrate that reducing RNA input from 250ng to 100ng required a sample size increase of 1.2 fold, and dropping to 10ng of input RNA required an increase of 2-3.5 fold. However, we were also able to demonstrate that the reduction in RNA input reduced sensitivity, i.e. we detected the expression of fewer genes, but without a significant impact on specificity, i.e. the genes we detected as differentially expressed were truly so (**Fig 2.8**).

Quality control

Performing rigorous quality control (QC) of experimental data during, and after, data collection and before starting biological analysis or interpretation can save time and effort in validation studies. The use of formal platform-specific QC steps has been a part of genomic and transcriptomic work since the technologies appeared, partly driven by their high cost. Sample QC allows researchers to verify that the samples are likely to generate “good” biological data. Methodological QC steps allow performance verification of key steps in the experimental process. Also, primary data QC steps allow only high-quality data to be processed for biological insight.

Quality control of nucleic acids

DNA and RNA samples used in sequencing experiments should be quality controlled before use against pre-defined experimental parameters. These may be ignored where samples of lower quality or quantity are all that is available for study. However, down-stream QC metrics may not be the same as those generated from high quality samples, and increased numbers of replicates may be required to compensate for the reduced quality or quantity⁵.

DNA is quality controlled by spectrophotometric, fluorimetric and gel electrophoretic methods. RNA QC may be more rigorous due to the high cost of microarray and RNA-seq methods relative to DNA PCR. Most RNA studies use the Agilent Bioanalyser (Agilent Technologies, USA) to generate an RNA Integrity Number (RIN), which provides a robust and non-subjective method for RNA QC. This has been shown to correlate highly with other methods¹²⁰, but allows a non-subjective selection of experimental samples. A RIN number of

>7 is generally used for microarray and RNA-seq experiments¹²¹.

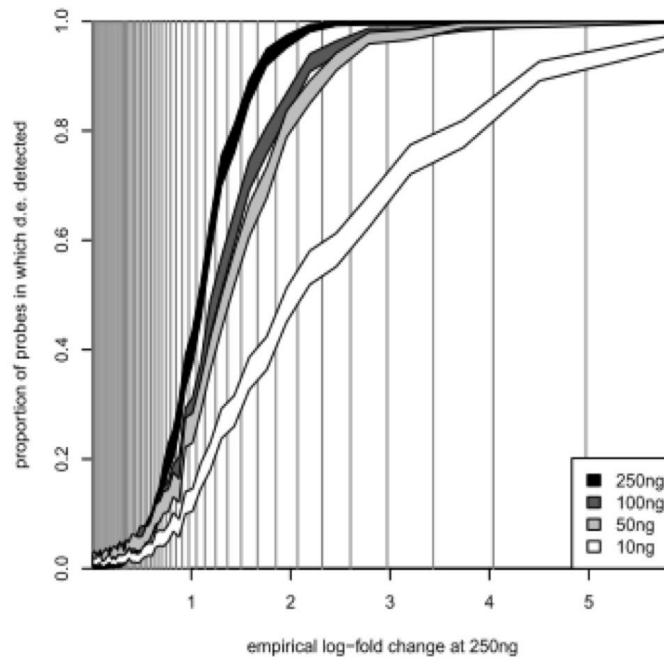
Quality control in next-generation sequencing

The high cost of NGS makes QC an important part of the experimental process. There are three main areas where QC can be applied during NGS: after library preparation, during sequencing, and before primary biological analysis. We demonstrated two of these QC steps in a ChIP-seq methods publication², providing guidance for assessing ChIP-seq library quality and visual QC of aligned ChIP-seq data (**Fig 2.9**).

Two NGS QC packages can be applied before primary biological analysis. FASTQC¹²² reports multiple QC metrics, including per base sequence quality score and GC content, duplication rate, etc. Multi-Genome Alignment¹² (MGA) presents NGS run data in visual and tabular formats to simplify QC assessment of run yield and quality. Both methods provide QC metrics that can be used to quickly identify common problems with NGS data and QC individual runs before primary biological analysis. However, results from these packages are context dependant; QC metrics for whole genome sequencing are different from those used for RNA-seq.

When designing the MGA tool we considered the metrics that users of NGS data might use to determine the quality of an experiment. NGS experiments are run in single or multiple sequencing lanes and most issues are contained within a lane. Many QC metrics are provided by Illumina, and by packages like FASTQC. We focused on those that were of high value and which were easily communicated visually, and designed the report to be used in the context of a very preliminary QC rather than an exhaustive analysis of run or sample quality. Therefore we present data for two primary metrics; sequencing yield and data quality in a lane-specific context and visualised as an Illumina flowcell is laid out.

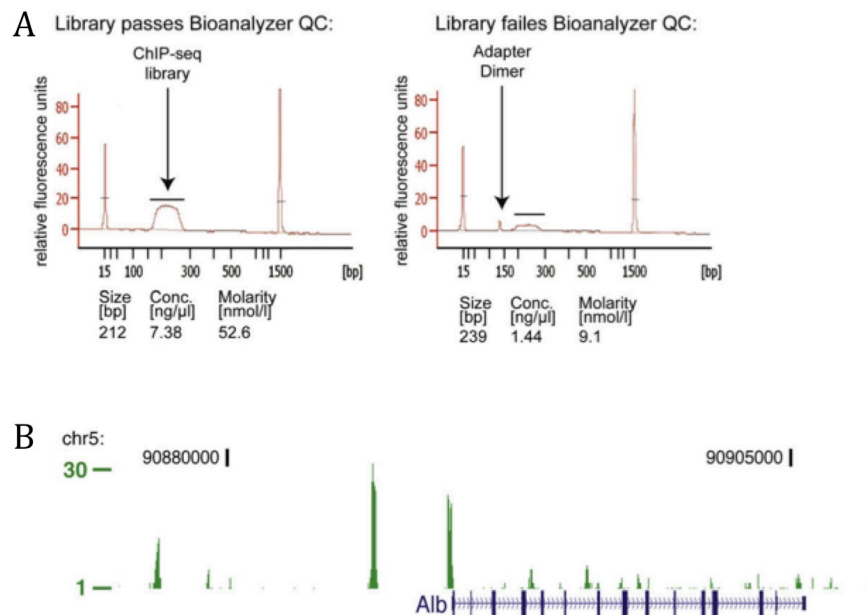
Fig2.8: RNA input affects DGE sensitivity



Sensitivity to detect differential gene expression decreases as RNA-input is reduced. 95% of DGE detected using a 250ng of input RNA can be detected at 100ng, and almost 50% at just 10ng. The increase in noise from using 10ng input RNA can be compensated for by increasing experimental replication. Reproduced from Lynch *et al* **BMC Genomics** 2010 ⁽⁵⁾

Fig2.9: ChIP-seq QC by capillary electrophoresis and analysis of a genome browser track

D. Schmidt et al./Methods 48 (2009) 240–248



A. Agilent bioanalyser traces for two ChIP-seq libraries. The left panel shows a successful library preparation, the right one that has failed with significant adapter-dimer contamination. **B.** A genome browser track for C/EBP alpha ChIP-seq at the Albumin locus showing strong and weak binding peaks.

Reproduced from Schmidt *et al* **Methods** 2009 ⁽²⁾

Chapter 3: The development of companion diagnostics

Introduction

A single mutational event can drive cancer, and act as a target for therapy, the canonical example being the Philadelphia chromosome and its targeted therapeutic - Imatinib^{123,124}. Over 95% of chronic myeloid leukemia (CML) patients have similar *BCR-ABL* fusions. The detection of the Philadelphia chromosome by qPCR¹²⁵ is used to monitor response to Imatinib therapy by detecting minimal residual disease. The development of Imatinib created the paradigm of targeted therapy where mainly cancer cells are killed. Because 95% of CML cases have the *BCR-ABL* mutation there is little need for stratification of patients. Other cancer driver mutations require a different paradigm, one where patients are selected for treatment with a specific targeted therapy based on their mutational status.

ERBB2 (synonyms include: *c-erbB 2*, *HER2*, *neu*) is the canonical driver gene in HER2+ breast cancer and is discussed in this chapter as a target for molecular tests that allow treatment with anti-*ERBB2* therapies. *ERBB2* is a powerful biomarker and is easily assayable at the DNA level, and for which several PCR based assays have been developed, including Jennings *et al*: **A differential PCR assay for the detection of c-erbB 2 amplification used in a prospective study of breast cancer**¹, one of the papers being submitted as part of this PhD by Publication. However after two-decades of research the preferred test is still immunohistochemistry (IHC).

In this chapter I will give an overview of the importance of *ERBB2*, its treatment with Trastuzumab and introduce the concept behind the paradigm of personalised-medicine with the use of molecular tests to select *ERBB2* amplified patients for treatment.

ERBB2:

The DNA copy number of *ERBB2* is amplified in 10-25% of breast cancers, this results in over-expression of the mRNA and functional HER2 protein increasing receptor mediated intracellular signaling, which drives aberrant cell proliferation and tumour growth. *ERBB2* is located on chromosome 17 (17q11.2-q12) and encodes the human epidermal growth factor receptor-2 (HER2) protein. This is a 185kDa transmembrane type 1 receptor tyrosine kinase¹²⁶, it is a member of the epidermal growth factor receptor family. Patients with detectable *ERBB2* amplification and overexpression have a significantly poorer prognosis than patients with normal *ERBB2* copy number; and increased likelihood of disease recurrence and reduced survival¹²⁷⁻¹²⁹. *ERBB2* has become an important therapeutic target in

breast cancer, and the monoclonal antibody Trastuzumab¹³⁰ (Genentech, USA) was approved for treatment of *ERBB2* amplified breast cancer in 1998. Recent data suggest it will also be an important target in other cancers; patients with advanced gastric cancer showed significantly improved overall survival when treated with trastuzumab in-addition to chemotherapy¹³¹, and the UK's National Institute for Health and Clinical Excellence (NICE) recommends trastuzumab as a possible treatment for *ERBB2* amplified metastatic gastric adenocarcinoma. Data from the International Cancer Genome Consortium (ICGC) pancreatic cancer sequencing project reported *ERBB2* amplification in 2% of cases leading to a clinical trial to assess treatment with trastuzumab⁶² (**Fig 3.1**).

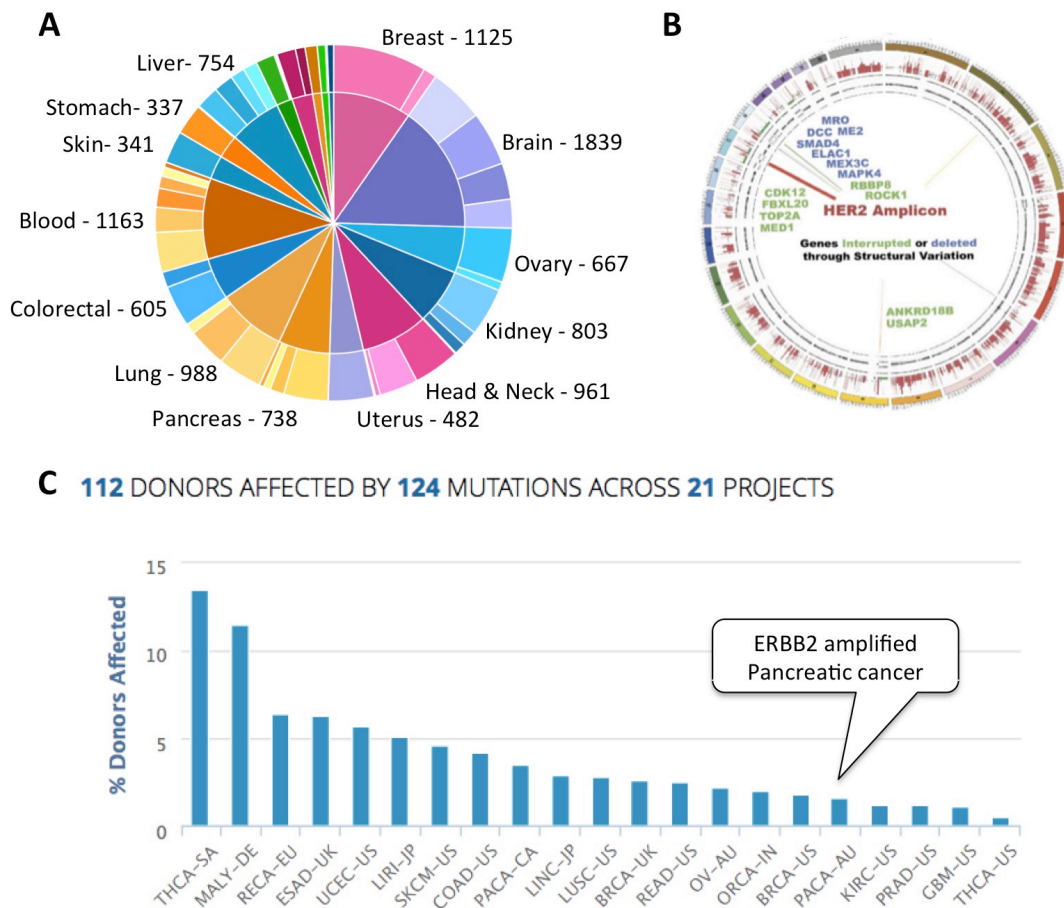
Trastuzumab binds with *ERBB2* blocking its dimerization and activation of the *RAS/MAPK* and *PI3K/AKT* pathways. Patients treated with trastuzumab achieve significantly increased progression-free survival, but only if they are *ERBB2* positive. Patients without *ERBB2* amplification receive no benefit from trastuzumab treatment, meaning patient selection by measuring *ERBB2* amplification status is key to successful treatment. As such *ERBB2* is a predictive biomarker for trastuzumab therapy, making *ERBB2* status both prognostic and predictive.

The reliability, sensitivity and specificity of *ERBB2* amplification testing are paramount because of the prognostic and predictive implications. A false-positive result could mean treatment with an expensive drug (currently \$40-100,000 per patient per year) giving no clinical benefit as well as possible side effects including flu-like symptoms, nausea, diarrhoea, and in rare cases cardiac dysfunction, whilst a false-negative result could prevent a patient from receiving a beneficial treatment. There are multiple methods to detect and measure *ERBB2* amplification status, including the trastuzumab/Herceptin companion diagnostic Hercep-Test (Dako, USA).

Measuring *ERBB2*:

ERBB2 amplification status can be determined from DNA, RNA or protein levels. A 2011 review of technologies for testing *ERBB2* amplification status in Breast Cancer¹³² concluded that although the previous decade had resulted in significant technical progress there was no consensus on a “best” test. It reported good overall correlation between different techniques, but did not evaluate sensitivity and specificity of the different *ERBB2* tests reviewed. A large multi-center study using a series of control cell lines with known *ERBB2* expression levels reported assay sensitivity for use in clinical trials¹³³. A more recent study used Tissue

Fig 3.1: Incidence of *ERBB2* amplification in Pancreatic cancer data from ICGC



A: ICGC has completed over 10,000 cancer genomes, colours denote the different cancers, further sub-divided by the national submissions. **B:** A circos plot for an *ERBB2* amplified Pancreatic cancer. Data is available through an open searchable archive. **C:** Analysis of pancreatic cancer genomes showed a 2% prevalence of *ERBB2* mutations suitable for Trastuzumab therapy.

Data from ICGC DATA RELEASE 15.1 <http://dcc.icgc.org>.

Microarrays (TMA) to test *ERBB2* with six different assays across 1210 breast cancers from six hospitals¹³⁴, and found *ERBB2* test sensitivity was 98% (94-100%) and specificity was 99% (97.9-100%). However laboratories continue to pick tests based on local knowledge and preference and the two main assays used for testing *ERBB2* amplification status in hospital laboratories today are immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH). Several PCR methods have been demonstrated but not widely adopted. The recent application of next-generation sequencing is making *ERBB2* copy-number analysis possible as part of a multi-target, and even multi-disease, assay; this final point will be discussed in chapter 6.

Measuring ERBB2 with IHC:

Immunohistochemistry (IHC) semi-quantitatively measures *ERBB2* protein levels and is the simplest technique for most laboratories to implement to assess *ERBB2* amplification. It is fast, cheap and with proper controls can be very sensitive. Tumour sections are incubated with an antibody to the *ERBB2* protein on the cell surface and then visualised. Two FDA-approved kits for IHC testing of *ERBB2* amplification are available: the Dako Hercep-Test (Dako, USA) and the Ventana Pathway *ERBB2/neu* test (Ventana, USA). Both quantitatively stratify *ERBB2* expression levels using the scores 0, 1, 2+, and 3+. Trastuzumab treatment would be prescribed based on a 3+ result (**Fig 2.1**).

The major limitation of IHC is the interpretation of results, which can be subjective. *ERBB2* staining can be affected by technical issues including tissue sectioning, fixation and processing, so quality control (QC) procedures, testing regimes and other quality assurance (QA) processes are commonly employed to improve results and may be mandatory in some settings. Results in the 0, 1 and +3 staining patients are generally considered unequivocal, however the +2 group would usually be referred for secondary- or reflex-testing using FISH.

Measuring ERBB2 with FISH:

The developers of Herceptin (Genentech, USA) stated that that FISH was their preferred method when selecting patients for trastuzumab therapy¹³⁵. They measured *ERBB2* amplification status by FISH in three separate clinical trials and discussed comparison to the “gold-standards” of solid matrix blotting, qPCR, or FISH, but not IHC. FISH was their preferred method as it had the highest precision. The FDA-approved *ERBB2* FISH assay is a multi-colour FISH assay with probes that bind to *ERBB2* and chromosome 17 (which carries *ERBB2*). The number of copies of the *ERBB2* gene can then be estimated by using the ratio of *ERBB2* to Chr17 fluorescence, samples with a *ERBB2* gene copy number ratio ≥ 2.2 are

considered as *ERBB2* amplified.

Measuring ERBB2 with microarrays:

Although microarrays can be used to molecularly classify breast cancer using the Mammaprint assay (Mammaprint®, Agendia, Irvine, CA, USA), they have had very limited clinical impact. We showed that it is possible to determine *ERBB2* amplification status using snpCGH and gene-expression arrays in a study of over 2000 breast cancers (Curtis *et al* 2012)⁷. In this we reported that 40% of breast cancers assessed by IHC had been recorded as *ERBB2* amplified (+1, +2 or +3, although only 7% were +2 or higher), that 22% were amplified as assessed by CNV microarray, and 12% by gene expression. The 22% amplified as assessed by CNA arrays compares well to figures of patients given Trastuzumab. Correlation to IHC was high with 92% of IHC *ERBB2* +3 scored samples recorded as amplified by CNA arrays; 56% of IHC *ERBB2* +2 scored samples were recorded as amplified by CNA arrays reaffirming that the decision to treat with Trastuzumab at this level requires careful consideration to avoid under- or over-treatment. We did not determine sensitivity or specificity for CNA, however given the subjectivity of IHC it may not be considered a gold-standard. The data from Curtis *et al* 2012⁷ was also used in a crowd-sourced bioinformatics competition designed to find the best model to predict survival^{136,137}. The winning algorithm used a “hallmarks of cancer” approach that used signatures of co-expressed genes corresponding to prognostic molecular events. It was consistent across random sub-sampling of experimental data and outperformed previous methods. Other microarray based prognostic tests are described in chapter 2.

Measuring ERBB2 with end-point PCR:

Development of a differential-PCR method for ERBB2 amplification status

In Jennings *et al* we demonstrated an end-point differential-PCR (d-PCR) based assay for *ERBB2* amplification status¹. d-PCR is a semi-quantitative method that co-amplifies a target gene with a reference control of known copy-number¹³⁸. Copy-number of the target gene is estimated from the ratio of intensities of the two separate PCR products when resolved as bands on a gel and analysed using automated gel-processing software.

Optimisation of the ERBB2 differential-PCR

This method required careful optimization. Samples were selected to have high-tumour content (>70%), something that has become a more obvious issue for genome-wide assays such as whole-genome sequencing (WGS), exomes or RNA-seq, and which will be discussed in later chapters. Normal DNA was extracted from blood for ten patients and 28 healthy

controls. Additionally two control cell lines were used; MCF7 – hemizygous for *ERBB2* and SKBR3 with 8 fold¹³⁸ over-expression of *ERBB2*, we also mixed SKBR3 and normal DNA to achieve a dilution series of *ERBB2* copy-number. All test samples were quality assessed using spectrophotometry and DNA concentration was normalized. Although d-PCR is relatively unaffected by starting DNA levels the final gel-based analysis can be affected by saturation of signal, consequently the d-PCR assay is more robust when DNA concentration is pre-defined as the range of PCR cycles where the reaction is still in the exponential phase. PCR primers were designed for *ERBB2* and *HBB* (beta-globin) to have non-complementary 3' ends and similar GC content and to amplify similar but easily resolvable fragment lengths of 91 and 110bp respectively. PCR was stopped after 35 cycles whilst still in the exponential phase of the reaction, as shown by empirical testing of cycle number with control samples. These tests showed that the sensitivity limit for detection was 25 cycles and the plateau phase was entered at 40 cycles, 35 cycles were chosen to maximize signal and assay sensitivity (**Fig 3.2**). All reactions and analyses were duplicated, alongside no-template-controls (NTCs). Products were resolved on Metaphor agarose (Cambrex Corporation, USA), a high-resolution agarose gel, stained with SYBR® Green (Molecular Probes Inc., USA) DNA stain and visualised using a CCD camera and Gelworks image processing software (Ultra Violet Products, USA). Amplification levels for the SKBR3:normal DNA dilution series gave linear results corresponding to the expected *ERBB2* copy number. 11 of 42 (26%) of samples showed *ERBB2* amplification, similar to other studies. The use of a carefully controlled PCR assay and an automated image analysis system meant that we achieved a robust and sensitive test. However the adoption of this test by other laboratories would have been complex, as it required both specialised staff and equipment.

Other PCR-based methods and citation of our differential-PCR test

Hubbard *et al*¹³⁹ completed a conceptually similar study to ours and obtained similar results. However they used an extremely complex radio-labelled d-PCR assay requiring a correction factor for the different numbers of [³²P]dCTP bases in each amplicon. An *et al*¹⁴⁰ developed fluorescent differential-PCR (fd-PCR), analysing products on an automated DNA sequencer. They characterised almost 200 FFPE samples and found *ERBB2* amplification in 26% cases. Their fd-PCR was an obvious improvement to the gel-staining methods, and the use of an automated sequencer reduced technical artefacts.

Johnson *et al*¹⁴¹ compared measurement of *ERBB2* DNA copy-number and protein expression levels using d-PCR and IHC using a dilution series control similar to ours, by mixing DNA from normal breast tissue and breast cancer cell lines to generate known levels of amplified

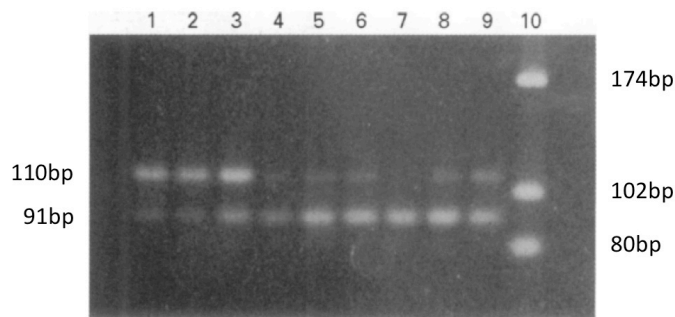
ERBB2. They showed saturation of d-PCR at 20 copies, and partial saturation of their immunohistochemical assay at a similar level. A 2003 review of *ERBB2* testing methods¹⁴² discussed the use of PCR-based methods, citing our paper, in the context of most laboratories using IHC and FISH. They concluded that at the time of their review PCR methods were still being evaluated and that, although they showed promise, the reality of pathology laboratories is the need to work with formalin-fixed paraffin-embedded (FFPE) samples and that the need to micro dissect to obtain high-quality results from PCR assays may limit their utility, although the more recent data suggest PCR assays are less affected by this than previously thought¹³².

The impact of Jennings et al on

The study was a well-designed test of the suitability of the d-PCR method for *ERBB2* amplification-status testing. It used a reproducible and objective method generating a numerical result for *ERBB2* amplification, and presented strategies for the design and use of controls to allow robust implementation of the test in other laboratories. We were able to draw conclusions about tumour biology in our patient series: the marker was associated with poor prognosis in our series of patients. We found that the mutation occurred before, and was maintained in, lymph node metastases. The work has been little cited primarily being due to several similar papers being published earlier, and also the development of real-time PCR making d-PCR obsolete as a quantification tool. IHC has also remained the dominant method for *ERBB2* testing due to its ease of use in clinical pathology laboratories as discussed earlier. The Jennings *et al* study was cited in the Johnson *et al*¹⁴¹ review discussed above. It has also been cited by Tsongalis and Reid¹⁴³ as a PCR example in their review of the value of *ERBB2* testing and the methods available. And by Naidu *et al* as a primary d-PCR reference in a paper examining FGF3 amplification in 440 breast cancers¹⁴⁴, and in their comparisons of IHC and d-PCR for c-MYC¹⁴⁵ and Cyclin D1¹⁴⁶ amplification and over-expression.

All the PCR studies discussed, including our own, were end-point assays. At the same time as we were completing our experiments a technique was being published that would make d-PCR quickly obsolete, real-time quantitative PCR⁷³ (qPCR). Commercial kits for the detection of *ERBB2* by qPCR were available as early as 2001⁸⁴, although IHC and FISH remain the standard methods. The success of phase III trastuzumab clinical trials was made

Fig 3.2: Differential PCR can be used to quantify *ERBB2* amplification



The ratio of intensities of the two bands produced during the differential PCR is used to determine *ERBB2* copy-number. The amplification products from *HBB* (110bp) and *ERBB2* (91bp) from three normal DNA controls (lanes 1-3), from three breast tumours with amplified *ERBB2* (lanes 4-6) and from the SKBR3 cell-line dilution series control equivalent to 8, 5 and 3 copies of *ERBB2I* (lanes 7,8 & 9). The 174bp, 102bp & 80bp bands of the molecular weight marker (lane 10).

Reproduced from Jennings *et al* **Mol. Pathol.** 1997 ⁽¹⁾

possible by the use of an IHC test to identify *ERBB2* amplified breast cancers¹³⁰. Recent research seems to point to next-generation sequencing based assays becoming the *de facto* standard for molecular-testing in personalised medicine^{8,9,147–150}.

Companion diagnostics

A companion diagnostic is a test, developed for use by any laboratory, that provides information that is essential for the safe and effective use of its corresponding therapeutic¹⁵¹, e.g. trastuzumab therapy is stratified by using a companion diagnostic to measure *ERBB2* amplification status. In September 2014 the FDA listed 19 approved companion diagnostic tests (**Fig 3.3**); 53% are for *ERBB2*, 16% EGFR, 11% BRAF, and just three other genes (KIT, KRAS and ALK) make up the other 20%¹⁵¹. 37%, 32% and 26% of tests use FISH/CISH, IHC or qPCR respectively, none currently use direct sequencing assays. In 2014 the FDA issued guidance on the identification and co-development of drugs and their companion diagnostic tests, aiming to improve the time taken to develop such tests. A laboratory developed test (LDT), by contrast to a companion diagnostic, is an in vitro diagnostic test developed for use in a single laboratory, but one which is not regulated in the same way as a companion diagnostic test. Many laboratories are developing NGS panels for use as LDTs and these are being developed to be used outside what might be considered the geographical area of the laboratory in which the test was developed¹⁴⁷. The FDA is currently developing regulation for LDTs. This is likely to have significant impact on the quality of tests, as well as the time taken to develop them.

In colorectal cancer the epidermal growth factor receptor (*EGFR*) is a therapeutic target with several anti-*EGFR* drugs available for prescription, however only a subset of patients respond to treatment. In 2009 the US Food and Drug Administration (FDA) updated advice on the use of these drugs requiring testing for activating *KRAS* mutations. In 2012 NICE approved Cetuximab (an anti-*EGFR* drug) in combination with chemotherapy as the recommended first-line treatment of metastatic colorectal cancer¹⁵², but in 2013 NICE suspended development of guidance on the clinical and cost effectiveness of using the different technologies and methods for *KRAS* mutation testing¹⁵³. *KRAS* mutations are found in just under half of patients and result in constitutive activation of the *RAS* signaling pathway. Mutations in *KRAS* are usually in exon 2, specifically codons 12 and 13. Furthermore mutations in *APC*, *MLH1*, *TP53*, *SMAD4*, *KRAS* and *BRAF* can be found in 10-100% of colorectal cancer patients¹⁵⁴ making screening for all 6 genes potentially useful, especially if this can be done at the same time as a *KRAS* test for deciding whether or not to proceed with anti-*EGFR* therapy¹⁵⁵.

In melanoma a specific mutation of *BRAF* (V600E valine-to-glutamic acid) is a target for treatment with Vemurafenib. The V600E mutation constitutively activates *BRAF* driving cell proliferation. This mutation occurs in half of melanoma patients and 5-10% of solid tumours, making screening for *BRAF* mutations in all cancer patients more likely as larger panel-based tests develop. *BRAF* mutational status is assed using a Vemurafenib companion diagnostic qPCR test approved by the FDA in 2013. NICE approved Vemurafenib for treating *BRAF* V600E mutation-positive unresectable or metastatic melanoma in 2012¹⁵⁶.

The need to match treatments to patients was apparent in 1997 when Jennings *et al* was published, however the idea that treatment would need to be personalised to such an extent as seems obvious today was almost unthinkable at the time. It was also not clear that a test would become so closely linked to a treatment to be considered a companion-diagnostic. Companion diagnostics are still not the norm, but are likely to become so as testing takes advantage of the variety of methods offered, a recent report suggested annual health care savings of \$604M if all colorectal cancer receive a genetic test for the *KRAS* mutation prior to treatment¹⁵⁷. However the slow adoption of improved methods over IHC for *ERBB2* amplification testing demonstrate how new tests need to be proven against current methods. The ability of NGS to assay multiple genomic loci, the exome or the cancer genome means it is likely to become an important tool for companion diagnostics. In the next chapter I will discuss how comparison of different technologies needs to be carefully considered to maximize the impact of research studies, and in chapter six I will discuss the emerging paradigm of multi-gene, even whole-exome, sequencing as a primary diagnostic test in cancer, using next generation sequencing.

Fig 3.3: List of FDA Companion Diagnostics

Drug Trade Name (Generic Name)	Target	Assay	Device Manufacturer
Xalkori (crizotinib)	<i>ALK</i>	FISH	Abbott Molecular
Mekinist (tramatenib)	<i>BRAF</i>	qPCR	bioMérieux
Zelboraf (vemurafenib)	<i>BRAF</i>	qPCR	Roche
Erbitux (cetuximab)	<i>EGFR</i>	IHC	Dako
Gilotrif (afatinib)	<i>EGFR</i>	qPCR	Qiagen
Tarceva (erlotinib)	<i>EGFR</i>	qPCR	Roche
Herceptin (trastuzumab)	<i>HER2</i>	FISH/CISH	Dako
Herceptin (trastuzumab)	<i>HER2</i>	IHC	Dako
Exjade (deferasirox)	Iron conc.	MRI	Resonance Health
Gleevec/Glivec (imatinib)	<i>KIT</i>	IHC	Dako
Erbitux (cetuximab)	<i>KRAS</i>	qPCR	Qiagen

A summary of the complete list of companion diagnostic approved by the US FDA. A companion diagnostic is defined as a device that provides information that is essential for the safe and effective use of a corresponding therapeutic product.

<http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm301431.htm>

Chapter 4: Comparing the analytical validity of platforms for genome and gene expression studies

Introduction:

It is important to develop research methods with high analytic validity; i.e. methods that are accurate in determining the presence or absence of single nucleotide variants (SNV) or mutations, copy-number variation (CNV) or differential gene expression (DGE). The sensitivity and specificity of tests is affected by the methods used at each stage: pre-analytical (e.g. sample collection), analytical (the test), and post analytical (data analysis and interpretation), all of which should be considered when choosing a method to use.

A common problem for researchers is choosing which method to use where several are available to accomplish a task; the same is equally true for clinicians, although the impact is arguably more significant. The primary question can be phrased as: “*do methods X and Y generate the same results when measuring the factor of interest?*” A systematic comparison of the different methods is the only way to reveal how similar they are. However it is difficult to perform a fair comparison outside of an idealised situation without bias since there are many variables to consider. In the comparison papers I have co-authored we specifically aimed to compare technologies with as little obvious bias as possible, to avoid comparing metrics which are obviously unfair and to highlight areas of bias and/or prejudice^{3,4}. In a review of micro-RNA technologies⁶ we discussed the imperfect nature of most comparison publications. We also drew attention to the fact that these publications are quickly out-dated. However none of these issues should deter researchers from undertaking comparison studies, especially as a pilot for a much larger piece of work^{7,158}. A 2012 editorial in the journal Nature Methods addressed the issues behind comparison testing¹⁵⁹. In this they highlight the choice a user faces: either expend resources on a formal comparison as a pilot to a larger experiment, or make a choice based on data already available that may not provide optimal results and hope this does not significantly affect the experiment.

Assessing the quality of comparisons:

Most users of comparison data do so as readers of comparison publications. They will be looking for data that support the use of one platform over another and must be critical in their evaluation of comparison publication. As such it is important for readers of comparison studies to be able to critically assess the results and conclusions presented yet only one paper has addressed this issue directly¹⁶⁰. In the clinical setting of this paper the authors state that new methods cannot be introduced without comparison to a reference method if one exists. However even if a gold-standard does exist it is unlikely to error-free and may be influenced by factors which lead to bias in assessing sensitivity and specificity¹⁶¹. Development of

reference standards and methods is being standardised by consortia like the Joint Committee for Traceability in Laboratory Medicine¹⁶², which brings stakeholders together and provides a framework for cooperation between very diverse groups. Laboratories developing *in-house* tests are using this framework to produce tests that are traceable to internationally recognised reference materials and/or reference methods. The National Institute for Health and Care Excellence (NICE) use a defined multiple technology appraisal process to review and assess the clinical and economic evidence behind potential new tests¹⁶³. A recent appraisal of testing options for epidermal growth factor receptor tyrosine kinase mutations in non-small-cell lung cancer reported that only 50% of tests were suitable for use in the NHS, and only when used in accredited laboratories that were also participating in an external quality assurance scheme¹⁶⁴. The US Food & Drug Administration (FDA) assesses new *in vitro* diagnostic tests using previously cleared assays or reference methods¹⁶⁵. Their primary recommendation is the use of the Type A Reference Method, which most comparison studies would refer to as a gold-standard i.e. one that has been thoroughly investigated and has been proven to be accurate and precise when used to analyse specific reference materials. If a Type A method is unavailable then a Type B Traceable Method can be used. This uses traceable and reproducible procedures and standards. The Type B study is a secondary recommendation because it requires initial agreement on, and development of, reference standards. In genome and transcriptome research several *de facto* standards are available: the Microarray Quality Control Consortium¹²¹ (MAQC) Brain and Universal Human reference RNA's, the External RNA Controls Consortium¹⁶⁶ RNA spike-in controls and the newly emerging NGS standards from the Genome in a Bottle consortium¹⁶⁷.

Bias in comparison studies

It is likely that a bias on the part of the researcher, experiment, or samples will mean one technology has an advantage where it might be disadvantaged in a different experimental situation. In the context of genomics research a particularly important issue to consider is that methods do not appear at the same time and that they, and their analysis methods, mature at different rates. This can make it impossible to determine if the one method is truly better than another or if it has just reached a particular point in its development cycle/lifetime that it performs best with the samples being tested.

Comparison papers should aim to reveal shortcomings in any assumptions made about samples, platforms and analysis methods by openly discussing them. The choice of biological samples to use in a comparison study and the methods for extraction, quantification, and

quality control can all affect results. Consequently many RNA comparison studies, but not all, use commercially available reference RNAs, however these may be poor substitutes for the experimental samples typically seen in a laboratory or clinical setting.

A semi-systematic review of mRNA microarray comparison studies

A literature review of microarray comparison studies was carried out to identify the scale and scope of comparisons to date with respect to mRNA gene expression microarrays. The review aimed to identify best practice in the use of controls and replication. The first microarray paper was published in Science in 1995⁹¹ at the time of writing there were over 61,000 microarray papers listed in PubMed. A search of PubMed from 1995 to 2013 found 352 papers using the search terms: *(microarray[Title]) AND comparison[Title]*, *((microarray[Title]) AND comparison[Title]) AND "gene expression"*, *((microarray[Title]) AND comparison[Title]) AND RNA[Title]*, *(microarray[Title]) AND compare[Title]*, *((microarray[Title]) AND compare[Title]) AND "gene expression"*, *((microarray[Title]) AND compare[Title]) AND RNA[Title]*, *(microarray[Title]) AND comparing[Title]*, *((microarray[Title]) AND comparing[Title]) AND "gene expression"[Title]*, *((microarray[Title]) AND comparing[Title]) AND RNA[Title]*. 97 duplicates were found and removed. 5 publications were not available as full text articles and were not included. 7 additional publications, found during the full text review, were included.

Inclusion and exclusion criteria

The titles and abstracts of 255 studies identified in the search were screened for those that compared multiple microarray platforms, or compared at least one microarray to an orthogonal technology. The screen was restricted to comparison studies directed at mRNA analysis. No other restrictions were placed on the type of microarray platforms used, the samples used or the study design. Most of the studies excluded were comparisons of sample groups, i.e. biological studies, rather than technologies (**Fig 4.1**).

Data extraction

A full text review of 26 screened publications was completed that extracted key details about the studies including: the number and type of technologies compared, were microarray methods compared directly or to an orthogonal technology or both, whether any validation method was used and if so what kind, whether and how studies were replicated, and whether

correlations were reported within and/or between technologies.

Results

The 26 included studies^{121,168–193} have been summarised in **Table 4.1**. They compared between one and six microarray platforms, generally with a single validation technology. The earliest comparisons suggested that microarray data from different platforms could not be directly compared, but almost all later comparison studies found good or excellent intra-, and good inter-platform correlations for some technologies. Later comparisons benefitted significantly from improved microarray manufacturing and processing as well as maturation in analysis methods, including normalisation methods¹⁷². Microarray processing was generally replicated, with 3-5 technical or biological replicates, although this was hard to determine from many studies. The studies often demonstrated higher correlations for commercial platforms, with Affymetrix microarrays scoring consistently highly and being used in 21 of the 26 studies reviewed. Later studies investigated intra- as well as inter-laboratory correlations in microarray results. One third of the comparison studies reviewed used commercially available reference materials. Two book chapters have also been published describing how to approach and interpret microarray comparison studies^{194,195}. They present a practical guide to the issues, detail common pitfalls to avoid and describe a framework for systematic comparison of mRNA expression profiling platforms.

The Microarray quality control (MAQC) consortium sets the standard for, and provides reference standards to use in, comparison studies

The publication of the MAQC papers was a landmark event for the field of transcriptomics^{121,186}. It involved over 100 scientists at 51 academic institutions as well as microarray technology providers and other stakeholders. Eleven microarray platforms averaging ~60 hybridisations per platform using two control RNA samples as reference standards, and calibration of microarray-observed DGE with other technologies including q-PCR showed that with careful experimental design different platforms can be highly comparable. The MAQC controls became the *de facto* standard for mRNA microarray studies. Ambion Brain RNA¹⁹⁶ is an equimolar pool of RNAs extracted and quality controlled from 23 individuals (13 male, 10 female). Stratagene's UHRR RNA¹⁹⁷ is an equimolar pool

Fig 4.1: The schema for a semi-systematic review of microarray comparison papers

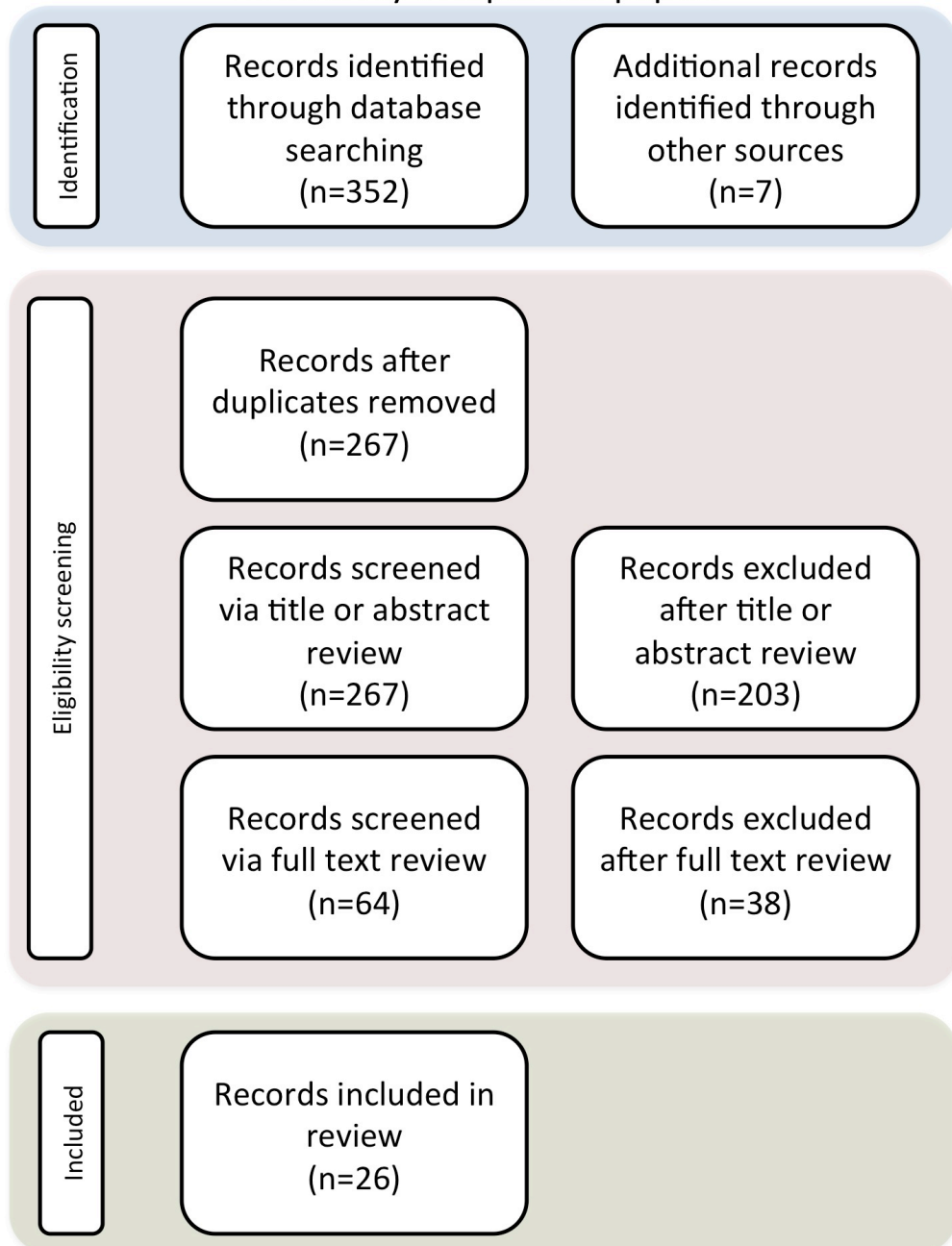


Table 4.1: mRNA comparison papers

Author	Year	Reference	Platforms compared						Validation method			Replication				Correlations reported			
			Number compared	Spotted array	Commercial	In-house	Oligonucleotide array	Allymetrix	Agilent	Other	Non-array	None	RT-qPCR	Other	Number	Type	Sample type	Within platform	Between platform
Kochapalli	2002	168	2		•				•				Undiarr	NA	Patient derived leukemic LGL cells versus normal PBMCs	Not reported	Poor correlation between arrays and qPCR		
Kuo	2001	169	2	•				•					0	The publication reported availability of small numbers of replicates but these were discarded before analysis	NC660 cell lines	Lack of replication made within platform comparison impossible	Very poor $r = 0.39$ and 0.46		
Piper	2002	170	1					•					16	Significance on arrays, 5 replicates on qPCR	Chemostat <i>S.cerevisiae</i>	95% correlation between laboratories			
Huen	2002	171	2	•				•					1-5	Significance on arrays, 5 replicates on qPCR	LFRTZ cells treated and untreated	Lack of replication made within platform comparison impossible	Modest correlation was reported for inter-laboratory replicate genes ($r = 0.72$)		
Barczak	2003	172	3					•					4-6	Technical replicates	K562 erythroleukemia cells and JH988 commercially available mouse tissues, in complex pools	Within-platform correlation was high ($r = 0.93$)	Reported high between-platform correlation ($r = 0.80$)		
Hyung-Jae	2003	173	2		•					SAGE			Undiarr	Technical replicates	Lack of replication made within platform comparison impossible	0.425-0.627 have a fairly good correlation in terms of absolute analyses and that the correlation is higher for genes with higher expression levels			
Regolina	2003	174	2					•		•			0	NA	D407 & ABE19 cell lines	Very high within-platform correlation: Allymetrix ($r = 0.92$), Atlas ($r = 0.94$)	Comparisons between the two platforms show a lack of agreement, no correlation reported		
Juran	2004	175	2					•					5-6	Biological and technical		Within platform variability was "small and purely a consequence of technical variability"	Between platform correlation was very high ($r = 0.94$) for genes classified as differentially-expressed on both platforms, but high false-positive rates		
Mah	2004	176	2	•				•					5	Biological replicates from patient samples	Colonic mucosa from colonoscopy screening	qPCR validation showed high correlation within Allymetrix ($r = 0.94$) and very high correlation with PDH arrays ($r = 0.93$), but low correlation between platforms ($r = 0.385$)	qPCR validation showed low correlation between platforms ($r = 0.385$)		
Parish	2004	177	2	•					•				Undiarr	Technical replicates	Commercially available mouse tissues, in complex pools	Not reported	Reported low correlations but no values		
Shippy	2004	178	2					•					5	Technical replicates	Commercial brain and pancreas RNA (Ambion)	Very high correlations shown in figures but no numerical values reported	High correlation $r = 0.79$ for genes called present in all replicates		
Woo	2004	179	3	•				•					4-6	Technical replicates	Mouse liver from A/J and C57BL/6J	Not reported	Overall concordance was good, the highest was between Allymetrix and long-oligo arrays, but no numerical values reported		
Yank	2004	180	6		•			•		•			3-6	Technical replicates	Matched mouse cell line and lung tissue	Allymetrix, Agilent and Codlink all had very high correlations ($r = 0.76$ to 0.93)	Allymetrix, Agilent and Codlink all had high correlations ($r = 0.72$ to 0.79)		
Trizary	2005	181	3	•				•					2-5	Duplicate within lab, 4,684.0 replicates between labs	AGS control cell lines and mice of these	Within-platform correlation was variable ($r = 0.40$ - 0.90)	Reported relatively good between-platform agreement		
Pyzdek	2005	182	3	•				•		•			4	Quadruplicate (not Allymetrix)	Arabidopsis	Correlation of signal intensities was high ($r = 0.86$ - 0.92)	For differentially-expressed genes correlations were as high as $r = 0.81$ (Allymetrix vs oligo)		
Schlegelmann	2005	183	2					•					6	Six biological	Six malignant HNSCC samples from patients	Very high within-platform correlation ($r = 0.98$)	Reported high between-platform correlations ($r = 0.61$ - 0.85)		
de Reghuis	2006	184	3					•		•			3	Technical replicates	Cell line, UHRR	Within platform correlations were great or than $r = 0.90$ for all platforms	Between platform correlations were high ($r = 0.72$ - 0.85)		
Kuo	2006	185	10	•				•		•			5	Technical (including separate laboratories)	Mouse	Very high correlations reported (0.71 - 0.95)	Very high correlations reported (0.63 to 0.92)		
MAQC consortium, Shi and Paterson papers	2006	121 & 185	11	•				•		•			3-10	Technical across time and space: 1329 microarrays hybridized at different times across 51 laboratories. These papers set the standard for comparison studies.	MAQC controls: Ambion Brain and Striatum UHRR - these became the de facto standards for gene expression analysis	Very high within-platform correlation ($r = 0.90$) for most platforms	High correlation seen across multiple sites and platforms ($r = 0.69$ - 0.87). Correlation with qPCR validation was very high ($r = 0.93$)		
Severgnini	2006	185	2		•					•			4	Technical replicates	MDA-MBP-231 cells	0.8-0.998, correlations	Inter-platform correlations were quite low (0.5 - 0.6)		
Chen	2007	185	3					•					1-2	Technical replicates	11 distinct Arabidopsis tissue types	The within-technology correlations across the tissues and mutants were greater than the between-technology correlations of identical RNA samples	Our analysis showed that the microarray and MPSS technologies did not correlate well on a quantitative basis for transcript abundance measurements		
Mouche	2008	185	2					•		•			5	Biological replicates from patient samples	Blood	Very high intra-platform correlation	Poor overall inter-platform correlation	0.51	
Hackley	2009	185	2	•				•					2	Technical replicates	Hipk2c cells	0.73-0.86	Not reported	Reported a statistically significant correlation for HER2 ($r = 0.67$) and uPA ($r = 0.07$) but not for PAI-1 ($r = 0.27$)	
Witzell	2010	185	2					•		•			2-96	large number of clinical samples, replication dependent on clinical features, eg: 53 HER2+, 35 HER2-, 5 HER2 unknown	Patient derived cancer samples	Not reported	Poor correlation of only 0.4-0.5	High correlations (>0.8)	
Su	2011	185	2							•			4	Biological replicates	Rat tissue	Not reported			
Xu	2013	185	2					•					3	Cell line replicates	Colon cancer cells	Microarray (0.94), RNA-seq (0.76)			

3,000 a

3-5 b

Table 4.1: mRNA comparison papers

26 mRNA microarray comparison papers were semi-systematically reviewed for this thesis. Over half used qPCR validation, but one-third used, or reported, no validation. The most widely tested platforms were Affymetrix then Agilent oligonucleotide arrays. The level of replication varied significantly with several publications not reporting any replication. Similarly with correlations, not all publications reported both intra- and inter-platform correlation metric and relied on qualitative descriptions for some comparisons.

of RNAs extracted and quality controlled from 10 human cell lines including: melanoma, liposarcoma, lymphoma, lymphoblastic leukaemia, myeloma and tumours of the breast, liver, cervix, testis and brain (glioblastoma). The commercial preparation and availability made these an excellent choice for microarray studies of differential gene expression. However as admixtures of RNA's this makes them very much less suitable for next-generation sequencing studies. The ideal control does not exist, as it will always need to have characteristics that match the context in which it is being used.

The MAQC study provided the first truly comprehensive technology comparison; it is the benchmark by which other comparisons should be judged.

Gold-standards

In developing a new test or assay it is advisable to identify control samples that can be used to assess both the sensitivity and specificity for that test. Samples with known results are often used, but much of the work in this thesis has had to be completed with samples of low-quality and quantity, both of these can and do affect the reliability or robustness of, and the ultimate sensitivity and specificity of an assay or test. A test result should be unequivocal; generating either true positive or true negative results, and reference samples that generate false positives and/or false negatives are suggestive that the test may have limited analytical sensitivity or specificity (**Fig 4.2**). As such both tests and a reference samples can be considered gold-standards. However accepted gold standards for one technology, e.g. qPCR for mRNA DGE, may not be applicable across the board. In the miRNA microarray comparison paper described below (Git *et al*⁴) we used a novel method to evaluate false-positive and false-negative rates after demonstrating that qPCR was not acceptable as a gold-standard for micro-RNA (miRNA) analysis. In developing the assays used in Murtaza *et al*⁹ and Forsheew *et al*⁸, both discussed in the next chapter, we used samples with known mutations to specifically assess both sensitivity and specificity.

Comparison papers published by the author

Curtis et al 2009: The pitfalls of platform comparison: DNA copy number array technologies assessed.

Considering bias in the design of Curtis et al 2009.

The Curtis *et al* copy-number variation (CNV) microarray comparison study³ was a pilot project, designed to inform the choice of platform for the largest study of breast cancer

molecular classification to date - METABRIC⁷. We acknowledged that it would be almost impossible to compare different technology platforms fairly. Although the final biological read-out may be the same, platforms have different methodological and technological characteristics. In designing, executing and analysing the study we tried to be aware of potential bias, avoiding it where possible. We excluded analytical tools that cannot be fairly applied to all technologies, and highlighted biases in the final publication. We had to consider comparison of one- and two-colour microarrays with markedly different DNA labeling, microarray hybridization and preliminary analysis methods. Although the intention was to choose the best performing system for copy-number; it was clear that the additional loss-of-heterozygosity (LOH) data provided by SNP microarrays (snpCGH) outweighed the higher per probe sensitivity of arrayCGH (aCGH) platforms and consequently we chose Affymetrix snpCGH over Agilent aCGH (the platform with the highest per-probe CNV sensitivity). We also discussed how cancer studies with only small amounts of and/or degraded nucleic acids can restrict researchers to a sub-optimal platform (as assessed by high-quality controls).

Selecting control samples for Curtis et al 2009.

During experimental design meetings preceding the study we discussed which controls to include allowing comparison of sensitivity and specificity between four different copy-number platforms. There were two conflicting approaches. Firstly to use standard well characterized HapMap¹⁹⁸ cell lines, with few copy-number aberrations. This approach makes it simple to determine sensitivity and specificity, but results may not be applicable to tumour samples. Secondly to use tumour samples and accept that determining sensitivity and specificity will be harder as there are likely to be many unknown mutations, and CNVs, of varying size. We chose to interrogate the issue with both types of control including the normal HapMap samples NA10851 (male) and a NA15510 (female), a series of tumour samples with and without matched normal DNA controls, and two cell-lines with known copy-number aberrations: MT3 and SUM159¹⁹⁹. We ruled out use of a well-defined X-copy cell line series^{100,200} as a model system to test sensitivity and specificity due to costs. The selection of controls was a lengthy process and delayed the experimental work, however given the final use of the data in choosing a platform for analysis of 2000-3000 tumour samples it was critical. It is important to point out that total cost was a limiting factor in the design of the study.

Results from Curtis et al 2009.

We were able to use known chromosomal copy-number events to interrogate and compare

Fig 4.2: Defining sensitivity and specificity

		Patient status	
		Condition positive	Condition negative
Test result	Test positive	True positive	False positive (Type I error)
	Test negative	False negative (Type II error)	True negative

$$\text{Sensitivity} = \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}} \quad \text{Specificity} = \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$$

False-positive and false-negative rates, sensitivity, specificity, precision, negative predictive value and accuracy can all be calculated from a matrix of test results using reference materials with known test results.

platforms. Deletions were most clearly detectable in the order of Agilent > Affymetrix > Illumina > Nimblegen. All four platforms identified single-copy gains of chromosomes 7 and 13, and single-copy loss of chromosome X in the MT3 cell-line. All four platforms identified the single-copy gains of the chromosome arm 5p in the SUM159 cell-line. Performance for high amplitude focal amplifications and sub chromosomal gains and losses was less clear. Three of the four platforms identified a complex 8q24 aberration in the SUM159 cell line (**Fig 4.3**), but some struggled with a 55Mb deletion. Small regions of gain/loss were assessed by analysing 79 PCR validated CNVs in normal HapMap controls. All four platforms had probes missing at multiple CNV locations, however the concordance amongst CNVs detected by all four platforms was high.

We assessed sensitivity and specificity of the four platforms using male and female samples and comparison of chromosomes X and Y to 13. This allowed us to test for copy-number changes 2:1 (Chr:X vs. 13) and 1:0 (Chr:13 vs. Y) at single-probe resolution (**Fig 4.4**). For copy-number 2:1 Agilent, Affymetrix and Illumina all showed similarly high sensitivity and specificity, whilst Nimblegen was poor. For copy-number 1:0 Agilent performed best, although Illumina outperformed it at very high specificity, Nimblegen demonstrated intermediate performance and Affymetrix was poor. Further investigation showed that this was due to a total lack of SNP probes on the Y chromosome for the Affymetrix array used in the study. Analysis of a public dataset for a newer Affymetrix microarray using a HapMap X chromosome titration data set²⁰¹, showed that the performance was unchanged for copy-number 2:1 analysis but dramatically improved for copy-number 1:0. These analyses demonstrate the challenge in comparing technologies that are rapidly improving, with the version of array having a dramatic affect on our results. The poor performance of the Affymetrix array could have resulted in its exclusion from our consideration of which platform to choose, but the availability of data from a newer version resulted in its being used in the METABRIC study.

Git et al 2010: Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression.

Git et al 2010 required a larger and more complex experimental design.

The miRNA platform comparison study we published was a pilot project for the largest study of breast cancer miRNA biology to date¹⁵⁸. The study design was larger in its scope than Curtis *et al* because the inherent biases of miRNA analysis make comparison of one or two platforms less justifiable. We compared nine miRNA differential expression platforms,

performing both intra- and inter-platform comparisons, making ours the largest miRNA comparison study currently published (**Table 4.2**). However it included only six of a possible nine platforms available at the time of publication. Previous miRNA comparison papers included fewer platforms and no current miRNA microarray platform has been included in all comparison studies as reviewed in Aldridge and Hadfield⁶.

Controls and validation used in Git et al 2010.

Prior to the miRNA comparison we compared the effect of RNA extraction methods on miRNA yield and quality. Cell lines were processed in duplicate across two extraction platforms, miRvana (Ambion, USA) and miRNeasy (Qiagen, Germany). mRNA Microarray analysis showed no significant differences in gene expression in the total RNA preparations.

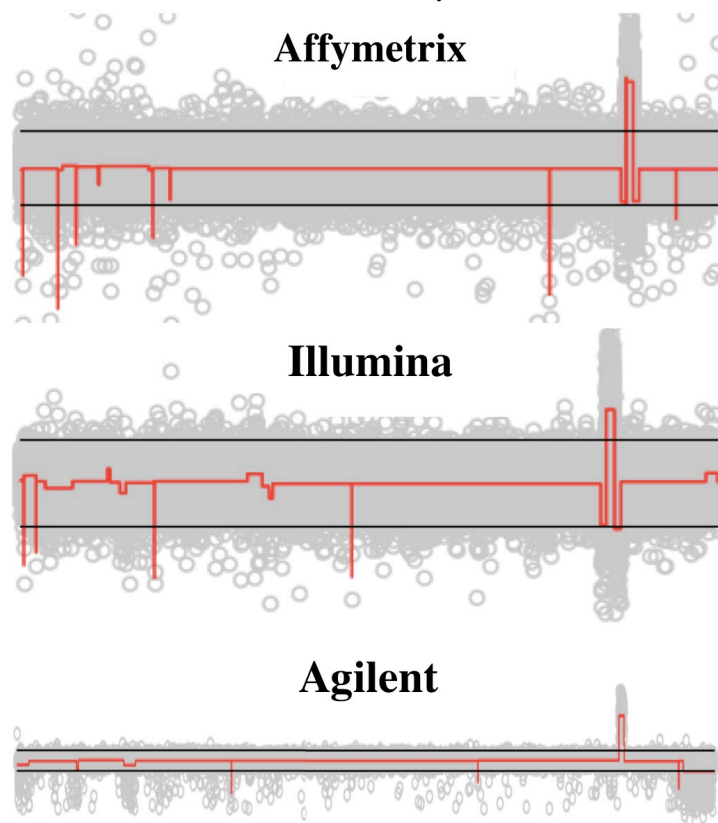
We used quadruplicate technical replication of three control samples: a pool of three commercially available normal breast total RNAs (normal), and RNA from breast cancer cell-lines PMC42 and MCF7²⁰². These were processed across six miRNA microarray platforms, and one next-generation sequencing method²⁰³. We compared platform utility and sensitivity and specificity for detecting DGE in miRNAs. Correlation between all platforms was high for a subset of miRNAs detected across all microarrays, that also had no predicted cross-hybridisation. NGS miRNA absolute read-counts were highly correlated to microarray hybridisation intensities.

Validation of nearly 90 miRNAs was completed using two qPCR methods (TaqMan and SYBR® Green), which showed very high correlations. Incorporation of this data into our analysis as a comparative method rather than as a gold-standard, led to a higher sensitivity across all platforms. This suggested that qPCR cannot be considered a gold-standard reference method for miRNA analysis in the same way as it can for mRNA DGE. However this analysis required the development of a novel algorithm to evaluate false-positive and false-negative rates for all methods in the absence of a reference method.

Conclusions

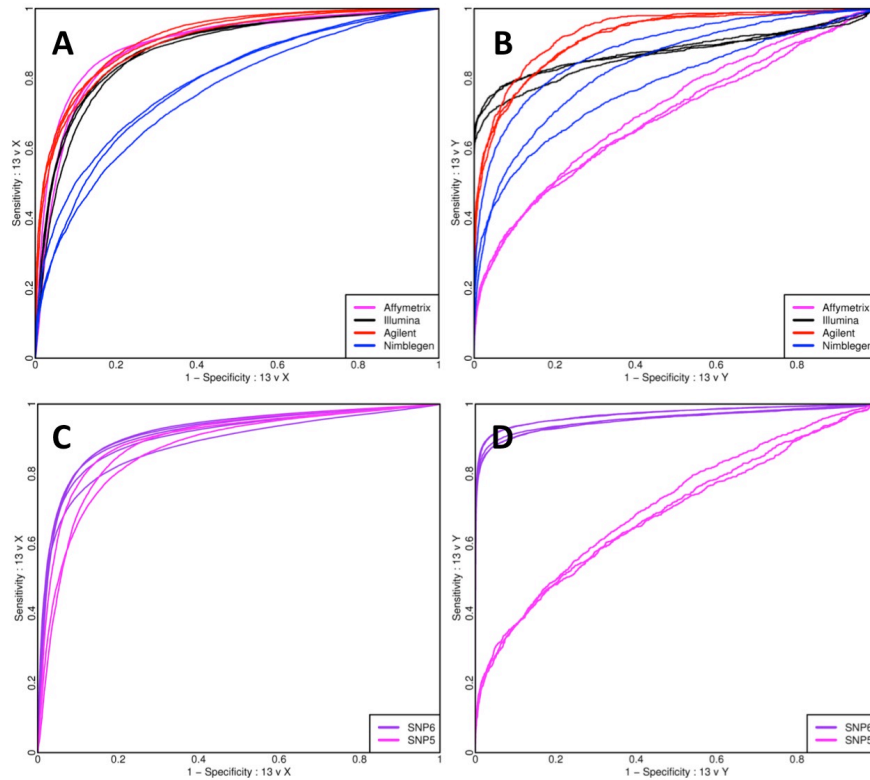
The development of genomic and transcriptomic methods continues at an incredible pace, and is one that authors of comparison papers cannot match. Until there is dominance of one system or method it is up to the reader of comparison studies to assess the bias within them and make a decision on which platform to use. This will be affected by the bias of the

Fig 4.3: Comparing performance of copy number microarrays



Copy-number variation at chromosome 8q24 in the SUM159 cell-line is detectable in three of four microarray platforms compared. The results shown are from a single replicate relative to a pooled normal in the Affymetrix and Illumina datasets. For the Agilent dataset, a single replicate from the dilution assay is shown. The Nimblegen data were not acceptable and are not shown.

Fig 4.4: Sensitivity and specificity of copy-number detection varies between different microarrays



Sensitivity and specificity of single-probes for detecting copy-number variation, from one to zero copies, is affected by the number of probes on the array. (A) for one-to-two copies Nimblegen performs poorly, (B) for two to one copies Affymetrix performs worst, but with significant variation between all four platforms. Affymetrix arrays with (SNP6) or without (SNP5) Y chromosome probes perform similarly for two to one (C) copy-number variation but are significantly better for one to zero (D) copies.

Reproduced from Curtis *et al* **BMC Genomics** 2009 ⁽³⁾

Table 4.2: miRNA comparison papers

	Number compared	RT-qPCR			Microarray Platforms										Sequencing	
		SYBR	TaqMan	ABI LDA	Af	Ag	Am	C	E	Il	In	L	T	Illumina	SOLiD	
Ach (2008)	2		●			●										
Chen (2009)	2		●									●				
Dreher (2010)	4		●		●				●		●					
Pradervand (2010)	5			●	●	●				●				●		
Sah (2010)	5				●	●	●		●	●						
Yauk (2010)	5			●		●			●		●	●				
Sato (2009)	6	●				●	●		●		●		●			
Baldwin (2009)	7			●	●	●			●	●				●	●	
Git (2010)	9	●	●			●	●	●	●	●	●			●		

Micro RNA microarray comparison papers were reviewed briefly in Aldridge & Hadfield 2012. Nine comparison papers were published in three years, nearly all used RT-qPCR validation as a “gold standard”, nearly all commercially available microarrays were tested and relatively new next-generation sequencing methods were also included.

ABI LDA: TaqMan low-density array cards, Af: Affymetrix, Ag: Agilent, Am: Ambion, C: Combimatrix, E: Exiqon, Il: Illumina, In: Invitrogen, L: LC sciences, T: Toray.

Reproduced from Aldridge & Hadfield **Methods** 2012 ⁽⁶⁾

researcher and his or her experiment and the need to balance time, precision, accuracy, cost, and sample type. Here I propose some simple rules that authors of comparison studies can follow when designing such experiments, and that readers can consider when assessing the quality of the published comparison.

1. Comparison study publications will ideally include some form of literature and/or technology review. It can be as important for readers to be aware of technologies excluded, as well as those included, when assessing bias in the study design.
2. Comparisons should be made:
 - i. to an accepted gold-standard reference method if one exists and/or
 - ii. with accepted reference materials if such exist, e.g. MAQC RNAs¹²¹
 - iii. or both of the above.
3. Test materials will ideally contain known and relevant test results that allow assessment of sensitivity and specificity.
4. Comparisons should include at least one comparative method and at least one reference method.
5. Replication with at least three biological or technical replicates should be considered the minimum.
6. Laboratory and analytical methods should be adequately described, such that any other laboratory could repeat them, and attention should be drawn to any known biases.
7. Raw data should be made publicly available where appropriate and made compliant with the relevant repository standards, e.g. MIAME²⁰⁴, MIQE⁷⁷ or MINSEQE²⁰⁵.

The lessons learned from previous comparison studies can be used when designing studies that make novel use of new technologies, and this will be discussed further in the next chapter.

Chapter 5: Application of genomic technologies in translational cancer research

Introduction

The Impact of next-generation sequencing on cancer biology

Next-generation sequencing (NGS) has had a significant impact on our understanding of the patterns of mutation present in cancers^{59,206,207}. Cancer was originally thought to arise from a single clone with a critical mass of somatic mutations gained in a linear manner. Recent studies have shown us how complex cancer evolution can be, with different mutation rates and types: kataegis⁵⁹ (localized hyper mutation) and chromothripsis (massive genomic rearrangement)²⁰⁸; the different mutational processes that underlie types or sub-types of cancer⁵⁹; and the impact of tumour evolution during cancer development and therapy^{9,61,64,209} (**Fig 5.1**). These are giving us new insights into cancer biology that can be used in the clinic to improve patient outcomes, but there are multiple challenges in translating these data and findings.

Analysis of cancer genomes using PCR-based and Sanger- sequencing methods initially used a candidate gene approach that is slow. In 2006 13,023 genes in breast and colorectal cancer were analysed using high-throughput Sanger sequencing of PCR amplicons⁶⁵. This revealed an average of 90 accumulated mutations in the cancers studied. A core set of 190 genes (11 per cancer) was mutated at high frequency, and many of these were newly reported as having functional consequences in cancer. However this study required PCR of 135,483 amplicons and sequencing of 3 million Sanger reads. The first NGS cancer genome was published in 2008²¹⁰ and today the International Cancer Genome Consortium⁵⁸ is sequencing the 50 most prevalent cancers in 500 individuals and recently started to release data²¹¹. This project is on a scale that would have required 3.4 billion Sanger sequencing reads just to interrogate the same regions as reported in 2006. The International Cancer Genome Consortium (ICGC) has presented a comprehensive picture of the spectrum of somatic mutations across, and within, different cancers (reviewed in Watson²⁰⁶). It has allowed a much clearer picture to emerge of which genes are the drivers of cancer progression, as well as revealing in much better detail the extent to which mutations are clustered in hot-spots within genes or spread randomly across them. It has also shown very clearly that cancer originating in the same tissue can have very different molecular pathologies, which are important in determining prognosis and treatment and may be diagnostic and/or prognostic. A common problem in the treatment of cancer is the heterogenous nature of many tumours^{64,212}, which makes molecular characterization difficult. This is particularly true when current technologies have a detection limit of around 1-2% mutant allele frequency^{8,9} in a background of normal DNA.

NGS cancer studies reveal evolutionary mechanisms

The elucidation of tumour evolutionary mechanisms by NGS was first reported in acute myeloid leukaemia patients²¹³. In each of 8 cases the tumour subpopulations detectable at relapse had a common origin from a founding clone. They also showed that while some sub-clones detectable at presentation were eradicated by initial chemotherapy, that the same treatment might contribute to relapse by driving the accumulation of new mutations in other sub-clones allowing therapeutic escape. This raises the possibility that by using our increasing understanding of the genetics of cancer we can prioritise the development of targeted therapies and reduce the use of broad-spectrum cytotoxic therapies that produce large numbers of novel mutations. Ultimately understanding heterogeneity and tumour evolution will improve our knowledge of cancer and its treatment⁶⁴, and will be vital in a personalised medicine context.

Estimating the amount of sequencing required:

The Lander/Waterman equation²¹⁴ is the most commonly used method for computing sequencing coverage and can be rearranged to compute the number of reads to sequence a genome, exome or amplicome (amplicon-panel) to a desired coverage. The general equation is: $C = LN / G$, which can be rewritten as $N = CG / L$ to determine the number of reads required (this is what is typically discussed when designing experiments).

C = redundancy of coverage, G is the haploid genome size, L is the sequence read length, and N is the number of sequence reads. In the examples below paired-end reads of 125bp from each end of a fragment are used, but these are converted to single 250bp reads for simplicity.

Human genome 30x coverage = $(30^{\text{fold}}) \times (3 \times 10^9 \text{ bp}) / (250\text{bp}^{\text{PE125}})$ and requires 360M reads.

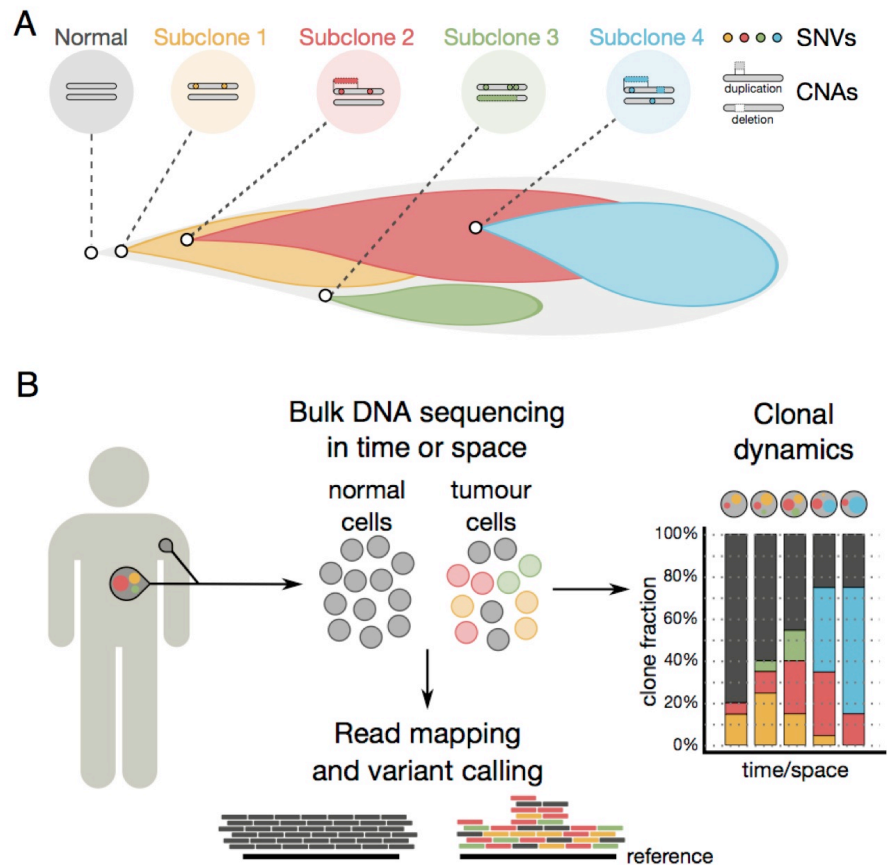
Human exome 50x coverage = $(50^{\text{fold}}) \times (1.5 \times 10^8 \text{ bp}) / (250\text{bp}^{\text{PE125}})$ and requires 30M reads.

Human amplicome ^(30x 250bp amplicons) 1000x coverage = $(1000^{\text{fold}}) \times (7.5 \times 10^4 \text{ bp}) / (250\text{bp}^{\text{PE125}})$ and requires 0.3M reads.

Choosing between amplicons, exomes and genomes

In 2014 Illumina announced the availability of the \$1000 genome, an astounding feat given that the first cancer genome cost around \$500,000 just six years earlier²¹⁰. The current gold standard is a PCR-free 30x-50x coverage Illumina sequenced Human genome. The ability to make next-generation sequencing libraries without PCR reduces the impact of GC bias in the sequencing process, and removes PCR duplicate reads. For most laboratories a genome is still

Fig 5.1: Reconstructing the clonal heterogeneity of cancer



(A) Schematic view of subclonal diversification. In this example, mutations in daughter cells of a single founder cell (left, grey) diverge into subclones (reflected by different colors) with distinct genomics features. (B) DNA sequencing of a tumor sample and its matched normal allow population structure to be inferred. The clonal dynamics can be determined using CNA and SNV information.

Reproduced from Fischer *et al* **Cell** 2014 ⁽⁶¹⁾

a significant undertaking as it requires high laboratory and bioinformatics resources. The first exome studies cost over \$1000 per sample but exomes in 2014 are around \$300-\$500, making them seem expensive when compared to the whole genome. However their main advantage is the significantly lower amount of sequencing data required; around 85% less than a genome at the same coverage (see above). Consequently many cancer studies are now using exome sequencing at significantly higher depth to investigate intra- and inter-tumoural heterogeneity. Exome analysis has been rapidly adopted by the clinical community for analysis of Mendelian diseases, where it can identify the casual variant in 25-35% of cases^{215–220}. By sequencing a trio of the proband and both parents, it is possible to identify causal *de novo* mutations (**Fig 5.2**). Inherited variation can be filtered and remaining variants can be screened for the 1-3 *de novo* mutations per exome predicted to be pathogenic^{215,221}. However the design of trio sequencing experiments needs to be carefully considered as sequencing artefacts can be impossible to distinguish from true *de novo* mutations. Results also need verification using an orthogonal method. The method is significantly faster than the “diagnostic odyssey” many patients go through with traditional single-gene testing. It also appears that the discovery of a casual variant, even if this has no treatment options, is a positive result for many patients and their families. A recently completed trio exome sequencing study of a patient with a life-threatening immunodeficiency by the author, resulted in the discovery of a likely causal *de novo* mutation (unpublished). The patient had undergone routine diagnostic immunological laboratory assessment that was uninformative. Extended diagnostics demonstrated an abnormal response to interferon, with failure to produce any detectable IL-12 and TNF-alpha, and a *de novo* dominant mutation was considered likely. Exome sequencing identified a *de novo* mutation in *NFKBIA* (Chr14:35,873,757 T>C) that leads to a loss of the I-kappa-B-alpha phosphorylation site abrogating NFkB signaling. The same mutation had been described earlier, and seven other *NFKBIA* mutations had been described in cases that phenocopied the patient. Sanger sequencing validated the mutation and a ‘single mismatch’ bone-marrow transplant appears to have been curative. Sequencing of amplicomes (NGS amplicon panels) requires even smaller amounts of data than genomes or exomes, making sequencing fast and very cost effective when run on a desktop sequencer like Illumina’s MiSeq (Illumina Inc., USA). Data can be generated and analysed in a diagnostically relevant turnaround time. A larger number of samples can bring statistical power to biological questions, however the depth of sequencing needs to fit the questions being asked, and an amplicome will miss variants found in an exome, which will miss variants found in a genome. Consideration of these factors during the experimental design is required.

Applying NGS technologies to circulating tumour DNA

Cell free DNA

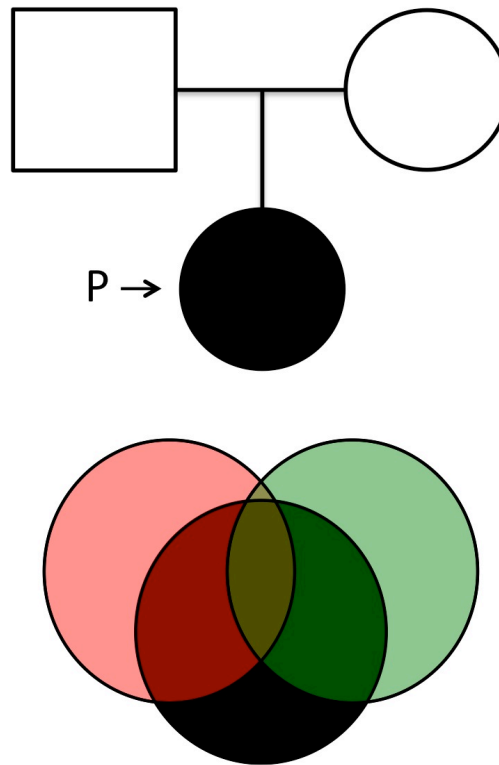
Circulating cell free tumour DNA levels (ctDNA) were shown to be elevated in cancer patients and to drop following treatment almost forty years ago²²². ctDNA was proposed as a noninvasive tool for cancer treatment in 2003²²³. Circulating cell free foetal DNA (cfDNA) is detectable in the plasma of pregnant mothers²²⁴ and was initially proposed as an alternative to circulating foetal cells for genetic screening²²⁵. It has subsequently been used to develop the revolutionary field of non-invasive pre-natal testing (NIPT): where cfDNA is sequenced to determine sex, aneuploidy, and sub-chromosomal abnormalities^{226–228}. The origins of cell-free DNA are still not clear although there appear to be two distinct mechanisms: lysosomal break-down of DNA from necrotic cells by macrophages²²⁹, or apoptotic break-down of DNA from hypoxic tumour cells²³⁰. cfDNA is fragmented to 140 to 170 base pairs, is present as thousands of amplifiable copies per milliliter of blood, but only a fraction is diagnostically relevant.

The concept of using ctDNA as a liquid biopsy in molecular pathology^{8,150,231–233} provides new opportunities to diagnose, prognose, monitor and manage cancer patients and their disease. The personalised genomics biomarkers used mean treatment can be tailored to the individual, and the method of sample collection will allow simple non-invasive longitudinal analysis. There is still much to be proven in how well ctDNA represents a patients whole tumour burden, but it almost certainly provides a better picture than single, and possibly multiple, biopsies. There are also standards that need to be developed for the protocols associated with collection and processing of blood for ctDNA analysis.

Development of ctDNA amplicon sequencing methods

PCR amplicon sequencing is a useful tool in cancer research; for whole exomes⁶⁵, or for single genes²³⁴, but in an era of whole genome sequencing the use of amplicons is in some danger of being over-looked as simply a validation tool to be combined with Sanger sequencing. However a relatively small number of amplicons can be highly informative: currently only nineteen companion diagnostic tests are FDA approved, and these cover just six genes: *ALK*, *BRAF*, *EGFR*, *ERBB2*, *KIT* and *KRAS*. NGS amplicon sequencing could target all of these in a single assay. Personalised medicine requires the development of tests that can predict the best treatment based on molecular evidence, and can be seen as an

Fig 5.2: Discovery of Mendelian disease causing *de novo* mutations using exome sequencing



Mutation detection relies on the fact that parental genomes are normal, and that there are few novel *de novo* mutations present in the patient. The pedigrees indicate the inheritance model underlying the strategy; empty symbols represent non-affected parents, filled symbols represent the affected proband. All samples are exome sequenced, and coloured circles indicate the genetic variants identified in the exomes of the mother (red), father (green) and proband (black).

extension of current tests like those developed for *ERBB2* and Herceptin (see chapter 3).

The development of ctDNA tagged-amplicon sequencing (TAm-seq) methods described in Forsheew *et al*⁸, presented a sensitive and novel method to detect and quantify tumour specific point mutations non-invasively and with high-sensitivity in patients with advanced disease. The use of ctDNA was an important advance over single biopsy methods, as a larger proportion of the tumour load can be assayed. It is also more convenient for patients, requiring only a blood sample to be taken, and should be cost-effective for the NHS and other health-care providers. The method required PCR amplification of just 5995 bases from the coding regions of *TP53* and *PTEN*, and selected genomics regions for *EGFR*, *BRAF*, *KRAS*, and *PIK3CA* with overlapping short amplicons, and covered 38% percent of point mutations in the COSMIC database^{8,235}. TAm-Seq noninvasively identified *TP53* mutations in 46 ovarian cancer samples. TAm-Seq identified the origin of metastatic relapse in a patient that presented with synchronous primary tumors (bowel and ovarian) 5 years earlier (**Fig 5.3**). TAm-seq also identified an *EGFR* mutation in plasma that was not detected in the initial biopsies. Lastly we used TAm-seq to track tumour dynamics over time using 10 coincident mutations detected in whole genome sequencing of a metastatic breast cancer patient (**Fig 5.4**). The 10 mutations followed the same pattern of initial decline in allele frequency, followed by an increase because of disease progression.

We also used TAm-seq as validation for somatic mutations discovered by exome-sequencing of aldosterone-producing adenomas which are the cause of 5% of adrenal hypertension cases¹¹. Tam-seq of *ATP1A1* and *CACNA1D* confirmed gain-of function mutations in these two genes important for the regulation of Na⁺ and Ca²⁺. The method was very efficient compared to Sanger sequencing validation.

The clinical utility of ctDNA and TAm-seq

Dawson et al¹⁵⁰ demonstrated how ctDNA can be integrated into clinical management of cancer patients, and how it may be a better choice of assay than cancer antigen biomarkers (e.g. CA125 in ovarian cancer) or circulating tumour cells (CTC). By comparing the performance of ctDNA, CA15-3, or CTC in 30 advanced breast cancer patients they were able to show detectable ctDNA in 97% of samples and reported somatic mutations in *TP53* and *PIK3CA*, with higher sensitivity than CA15-3 or CTC. They also reported that the absolute level of ctDNA corresponded to treatment response and survival.

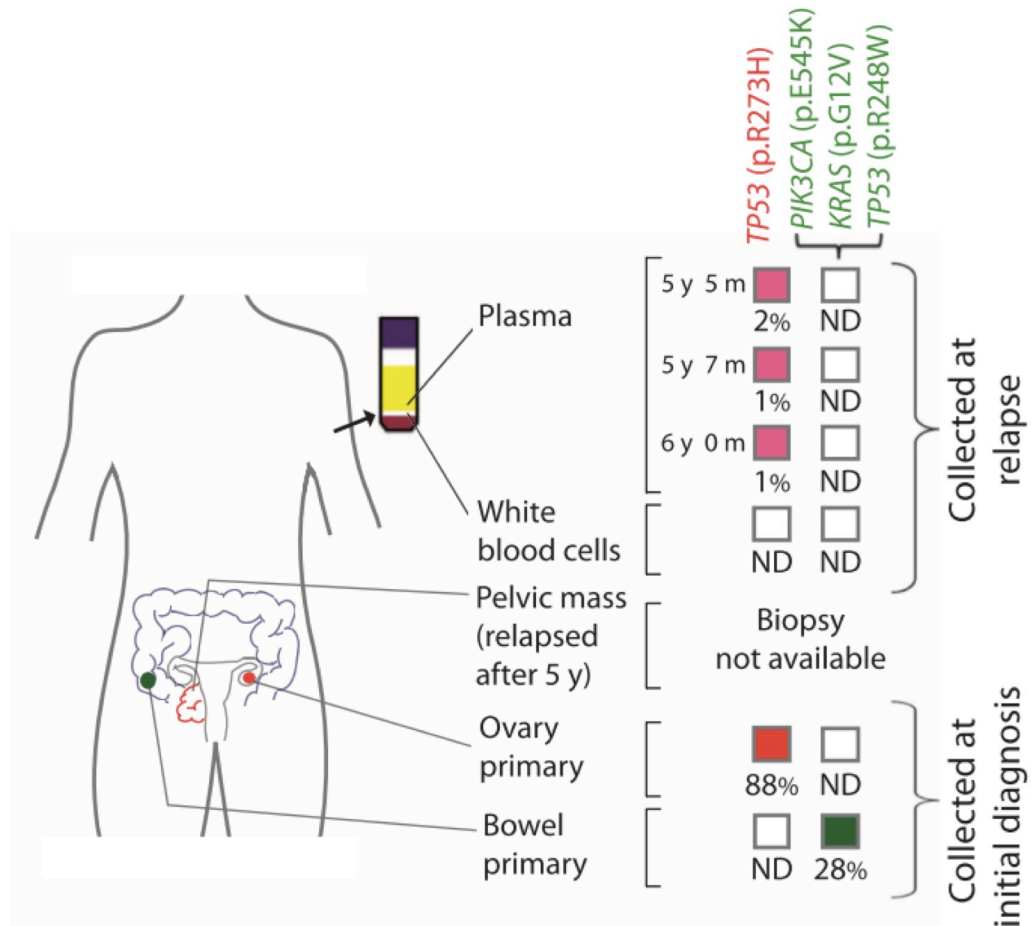
Other methods for assaying ctDNA do not compare well to TAm-seq

Other methods have been developed to assay mutations in ctDNA but these have limitations that TAm-seq does not. BEAMING^{229,236}, and digital PCR¹⁴⁹ use emulsion PCR or microfluidics and locus-specific assays to very accurately detect and quantify specific mutations, but cannot easily be used for genes without mutational hotspots. Although these are the most sensitive methods available, the requirement for custom primers or probes to be synthesized for a specific known mutation in a single patient's tumor, makes it impractical to use these assays for cancer screening²³⁵. TAm-Seq is not allele-specific and allows robust and accurate measurement of tumour-specific DNA across sizable genomic regions in blood plasma in a high-throughout and cost-effective manner. Its development opened up possibilities for large-scale studies to investigate the clinical utility of ctDNA as a non-invasive monitoring tool for cancer management.

Development of ctDNA exome sequencing methods

The successes of TAm-seq and the use of ctDNA to detect and quantify circulating tumour mutations led to the development of the whole exome sequencing methods presented in Murtaza *et al*⁹. We aimed to develop a method that would be less affected by, indeed would allow analysis of, tumour evolution. TAm-seq can be affected by allele drop-out; where a mutation detected at presentation is no longer present at sufficient frequency to be detectable, or has been lost altogether. This can happen as tumours evolve under therapy, but the much larger number of mutations found using whole exome methods should be robust to this. The large number of potential variants assayed would also increase the possibility to monitor tumour evolution by detecting changes in allele frequencies due to tumor evolution. We reported the use of a modified exome capture protocol to allow the use of as little as 2.3ng of DNA (mean 13ng, min 2ng, max 40ng), equivalent to around 10% of ctDNA extracted from 2.0-2.2ml of whole blood (**Table 5.1**). This ctDNA was used as the input for a non-standard Illumina library preparation using a proprietary technology called ThruPLEX (Rubicon Genomics, USA). The ThruPLEX kit uses a modified Illumina method, with hairpin rather than Y-shaped adapters, optimised chemistry and bead-based size-selection and cleanup rather than gel electrophoresis. Three to five barcoded sequencing libraries were prepared for each case before pre-capture pooling for exome library production. Only 4-20% of the DNA extracted from total blood plasma was used for library preparation, using more may allow for an increase in analytical sensitivity and should be a significant advantage in the clinic, especially when comparing repeat blood sampling to repeat biopsy.

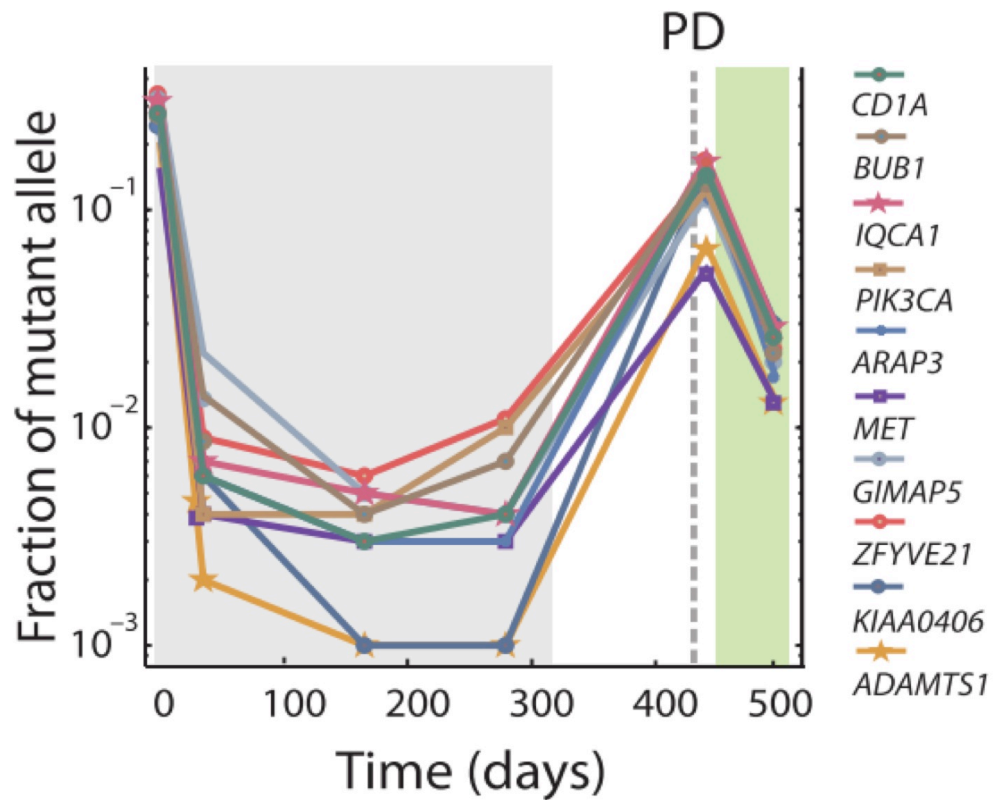
Fig 5.3: Confirming site of tumour origin using ctDNA



Retrospective analysis of samples from synchronous primary tumors (bowel and ovarian) collected at the time of initial surgery and three plasma samples collected at relapse. A TP53 mutation was identified in the ovarian primary (red box), mutations in PIK3CA, KRAS, and TP53 were identified the bowel primary (green box). At relapse ctDNA sequencing identified only the TP53 from the ovarian primary tumor (p.R273H).

Reproduced from Forsheew *et al* **Sci Trans Med** 2012 ⁽⁸⁾

Fig 5.4: Tracking tumour dynamics using ctDNA



Ten mutations identified from whole genome sequencing were assayed using TAM-Seq. Short amplicons (<120 bp) were designed to the mutation loci. Serial plasma samples collected before and after treatment were analysed to determine the status of mutant allele dynamics over time and during treatment.

Reproduced from Forsheew *et al* **Sci Trans Med** 2012 ⁽⁸⁾

Table 5.1: Preparation of ctDNA exome libraries is possible with variable inputs of ctDNA

Case	Cancer type	Sample	DNA amount (ng)*	Percentage of extracted DNA used	Percentage of plasma used*	Number of reads (millions)	Unique coverage (fold)
1	Breast	E1	23	4%	4%	260	147
		E2	2.3	4%	4%	179	47
		E3	9.7	4%	4%	236	160
2	Breast	E1	18	20%	2%	115	77
		E2	10	20%	2%	126	80
		E3	23	6%	2%	227	140
3	Ovarian	E1	6	7%	7%	152	49
		E2	24	7%	7%	278	131
		E3	9	7%	7%	223	70
4	Ovarian	E1	26	14%	14%	150	74
		E2	12	14%	14%	173	73
		E3	40	14%	14%	129	84
5	Ovarian	E1	5	14%	14%	156	49
		E2	8	14%	14%	147	60
		E3	4	14%	14%	137	40
		E4	10	14%	14%	329	84
		E5	14	14%	14%	227	96
6	Lung	E1	5.6	20%	20%	169	44
		E2	6.2	20%	20%	133	31

*Based on quantification of individual loci by digital PCR.

**The effective volume of plasma (only a fraction of the eluted DNA was used), followed by the original plasma volume from which DNA was extracted (in parenthesis)

Details for the circulating tumour DNA exome library preparation. Replicate exome libraries were prepared for all tumour samples using variable amounts of DNA (column 4), and percentages of plasma (column 6) available. Sequence coverage of protein-coding genes ranged from 31-fold to 160-fold.

Reproduced from Murtaza *et al* **Nature** 2013 ⁽⁹⁾

Exome sequencing allele frequencies correlated well with TAm-seq and digital PCR ($r = 0.71$). 60% of mutations detected in either patients 1 or 4 were found in both plasma and metastatic biopsies and mutant allele frequency were high ($r = 0.7$) demonstrating high specificity. A panel of 364 non-synonymous mutations were detected with high confidence and included previously reported cancer genes and genes reported as being involved in treatment resistance or disease progression:

- Case1 (breast cancer): following paclitaxel treatment an activating *PIK3CA* mutation was detected. Mutation of *PIK3CA* has been reported as linked to paclitaxel resistance in mammary epithelial cells²³⁷.
- Case 4 (ovarian cancer): following treatment with cisplatin an increased abundance of a truncating mutation in, and LOH around, *RBI* were detected. Loss of *RBI* has been linked to chemotherapy response²³⁸.
- Case 6 (lung cancer): following treatment with gefitinib an activating mutation in *EGFR* (T790M - substitution of methionine for threonine at position 790) was detected. The T790M mutation has been linked to acquired resistance to gefitinib therapy²³⁹ and the FDA approved afatinib for lung cancer in 2013²⁴⁰.
- Case 2 (ER+ ERBB2+ breast cancer): was particularly interesting as it was possible to detect emergence of resistance in canonical resistance genes after two rounds of therapy with different therapeutic agents. Following treatment with tamoxifen and trastuzumab a nonsense mutation in *MED1*, was detected, which has been associated with tamoxifen resistance²⁴¹. Following secondary treatment with lapatinib and capecitabine, a splicing mutation in *GAS6* was detected, this has been linked to resistance to lapatinib in ER-positive, ERBB2-positive breast cancer cell lines²⁴².

Sensitivity and specificity of ctDNA analysis:

In TAm-seq the generation of around 18,000 single nucleotide variants introduces the potential to generate a high-level of false-positive results. We used several strategies to control for this. Sample preparation was performed in duplicate, and variants were only recorded if they appeared in both replicates. 38 of 40 variants of allele-frequency >2%, as assessed by digital PCR, were detected and quantified by TAm-seq giving an assay sensitivity of >95%. However although we reported detection of *TP53* mutations in over 50% of high-grade serous ovarian cancers, previous reports confirmed the almost universal mutation of *TP53*, suggesting sensitivity might not be as high as first thought. We did not formally assess

sensitivity or specificity in Murtaza *et al* as the number of subjects was small. However we tried to maximise analytical sensitivity and carried out analysis on patients and samples with time points selected for high mutant-allele fraction.

The clinical utility of ctDNA

Neither of the studies discussed above was able to answer the question as to what the utility of ctDNA would be across varying cancers. A landmark study by Bettegowda *et al*²³³ reported analysis of 18 cancers in 640 patients. ctDNA was detectable in >75% of patients with advanced pancreatic, ovarian, colorectal, bladder, gastroesophageal, breast, melanoma, hepatocellular, and head and neck cancers; but in less than 50% of primary brain, renal, prostate, or thyroid cancers. ctDNA was detectable in 50-73% of patients with localized tumors, suggesting it may not be restricted to patients with advanced or metastatic disease. ctDNA was often present in patients without detectable circulating tumor cells, suggesting that these two biomarkers are distinct entities.

The study determined sensitivity and specificity of clinically relevant KRAS gene mutations as 87.2% and 99.2% respectively. False-negatives were generally associated with lower tumour burden and most likely lower ctDNA levels. ctDNA was also used to detect mutations key to EGFR therapy resistance, guiding treatment decisions in colorectal cancer.

Current status of clinical testing and adoption of NGS assays

Today cancer is still largely diagnosed by histological analysis of tumour cells taken as biopsies and/or after surgical resection. Molecular tests are still limited to germline hereditary disease with gene-by-gene sequencing, e.g. *BRCA1* and *BRCA2*. The impact that sequencing based assays can have in determining treatment, the discovery of tens or hundreds of mutations and the relative ease with which sequence data can be generated today are all leading to molecular tests, or molecular pathology, as being the likely future standard. Much of this is likely to be completed using next-generation sequencing of amplicons, exomes or genomes.

NEQAS has external quality assurance schemes²⁴³ in place or in development for - lung cancer: ALK rearrangement, and KRAS, BRAF and PIK3CA mutation screening; Colorectal cancer: BRAF and PIK3CA mutation screening; Melanoma: NRAS and KIT mutation screening; and Gastro-intestinal stromal tumours: KIT and PDGFRA mutation screening.

NEQAS are also piloting external quality assurance schemes for next-generation sequencing, using a reference DNA sample²⁴⁴. The US-FDA has approved three drug treatments that require a sequence based companion diagnostic: vemurafenib and *BRAF* for melanoma, cetuximab and *KRAS* for colorectal cancer and crizotinib and *EML-ALK*, *EGFR* for lung cancer¹⁰. Many more are currently in the approvals pipeline, and many of these are NGS multi-gene panel tests.

The dramatic impact that circulating cell free DNA is having in cancer, and NIPT, suggests that NGS-based molecular tests using ctDNA are likely to become a clinical standard.

Chapter 6: Discussion

That cancer is a heterogeneous disease is undisputed, but the level of heterogeneity and the evolution of cancer during treatment have only recently been measurable at the genome-wide level. Early studies are revolutionizing our understanding of cancer and highlighting therapeutic opportunities, particularly with targeted therapies; but only where a companion diagnostic exists that can be deployed in the novel patient population with sufficient specificity and sensitivity. There is a clear separation between the use of tests that detect the presence of a cancer driver mutation and those that quantitatively measure the changes in those same mutations. The ongoing International Cancer Genome Consortium (ICGC) projects will add significantly to our understanding of the key driver mutations for different cancers, and the prevalence of those mutations in the different diseases. Translating these technological developments and new biological insights such that they can be used in the clinic is challenging. The next decade is likely to see a revolution in the personalisation of both cancer, and non-cancer medicine.

Oncogenes and tumour suppressor genes can be analysed to determine therapy

Oncogene mutations can be found at very high frequency in some disease populations. If there are targeted treatments for these high frequency oncogenes then pre-screening of patients is not necessary, e.g. in the case of *BCR-ABL* positive chronic myeloid leukaemia (CML) where the treatments of choice are imatinib (Gleevec, Novartis) and second generation tyrosine kinase inhibitors, for nearly all patients. However these cases are rare. Most targeted agents are available for oncogenes present at lower-frequency in a patient population, and are only likely to show clinical benefit in a minority of patients. These patients must be selected from the general population with some kind of clinical test, often an IHC, FISH or PCR, but more likely in the future a genome-sequence based test. The comparison of CML to *ERBB2* amplified breast cancer in chapter 3 introduced the concept of using mutational status to stratify patients for treatments with targeted therapies; e.g. trastuzumab in the case of *ERBB2* amplified breast cancer. Although the current test for *ERBB2* amplification is IHC, other methods including differential-PCR¹, qPCR and microarrays have all been investigated. We showed in the METABRIC study⁷ that microarray analysis of breast cancers improved the sub-classification of this disease (**Fig 6.1**). We also showed that the microarray data could be used to infer *ERBB2* copy number status and that this correlated well with IHC data. In an analysis of ctDNA using Tam-Seq⁸, and later using ctDNA-exomes⁹ we showed that next-generation sequencing was a sensitive tool to detect and quantitate mutant alleles in

heterogenous cancer. The ctDNA exome data also revealed the evolution of resistance mutations, some of which were present in the initial, presumably heterogeneous, tumor. Understanding the mutational status of heterogeneous tumours is likely to improve patient treatment. Although the tools for next-generation based companion diagnostics are still in development, the work presented in this thesis demonstrates how far we have come in the last fifteen years. At a technological perspective we now have sensitive and specific technologies and assays for detecting and quantitating mutations in cancer samples. The focus for the future will be on the translation of these from ‘research use only’ to the clinic thus enabling personalised medicine.

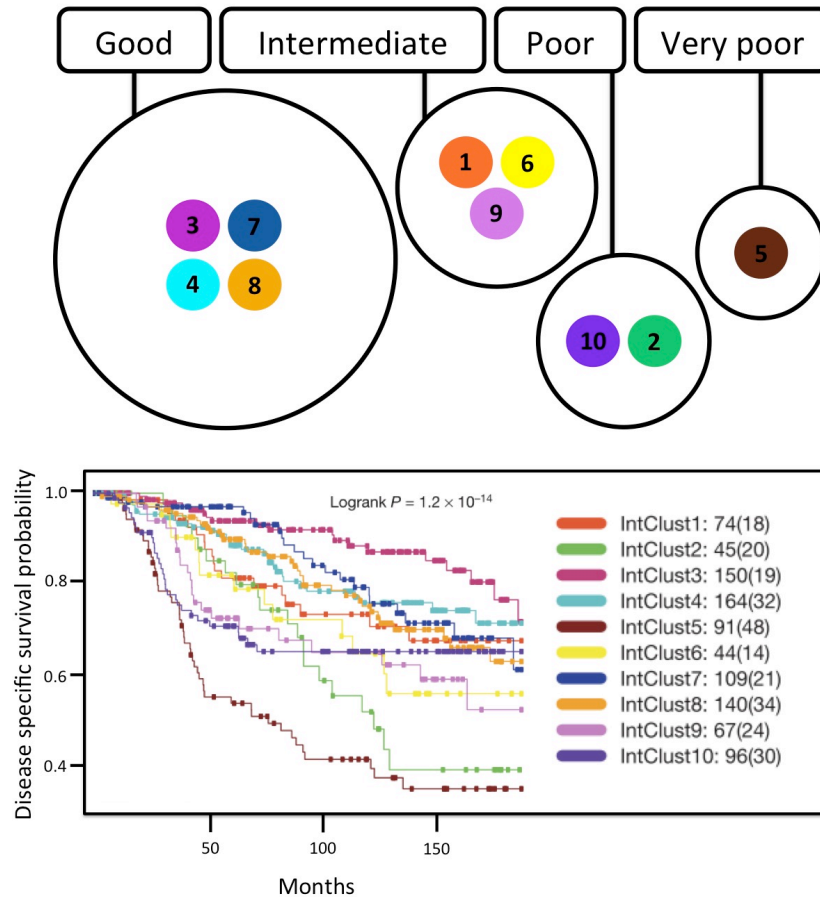
Personalised medicine and companion diagnostics are still in their infancy

The revolution that is being termed personalised medicine began with trastuzumab²⁴⁵ as discussed in chapter 3. The FDA approved Trastuzumab in 1998 for the treatment of *ERBB2*-amplified breast cancer, but this approval was conditional on the use of the companion diagnostic to determine *ERBB2* amplification status to stratify patients for treatment. The US FDA has approved 13 companion diagnostics, nearly all of which are based on just six gene mutations important in the treatment of cancers (**Fig 3.3**)¹⁵¹. Currently over half of FDA approved companion diagnostics are for stratification of *ERBB2* amplified breast or gastric cancer patients. Other tests include: testing for *BRAF* mutations identifies malignant melanoma patients that will respond to vemurafinib²⁴⁶, and testing for *EML4:ALK* fusions identifies non-small cell lung cancer patients that will respond to crizotinib²⁴⁷. These tests could be run as a combined NGS gene panel; but at the present time, and when compared to conventional tests, the perceived cost and complexity of these tests is considered too high. The cost of next-generation sequencing however, has dropped significantly; the HGP genome cost approximately \$300 million to sequence²⁴⁸, the Watson genome around \$1 million²⁷, and the current cost is \$1000 per genome. To put this in perspective the cost of health-care in the UK is around £2000 per person per year²⁴⁹. A \$1000 genome (if an individual is sequenced once, and at birth) need only produce an £8 per year saving to be cost-effective: about the same as a single prescription. We estimated the costs of TAm-Seq analysis to be around £25-50 per patient, for a seven-gene panel⁸. It should be cost-effective today to merge all six genes approved in FDA companion diagnostic tests, into a single NGS panel.

Patient response to therapy is heterogenous

The treatment of cancer has changed remarkably, and survival rates have doubled, in the last

Fig 6.1: Redefining breast cancer



The molecular classification of breast cancers using whole genome copy-number and gene expression microarrays redefined breast cancer as having ten molecularly distinct breast cancer sub-types each show differing prognosis and survival rates. 57% of patients have good prognosis, 19% intermediate, 14% have poor and only 5% have very poor prognosis.

Data from Curtis *et al* **Nature** 2012 ⁽⁷⁾

forty years mainly due to earlier detection and better treatments⁵⁵. Patient stratification has allowed drugs like trastuzumab to become the backbone of therapy for their respective diseases when stratified by a molecular test. Many breast cancer patients may be effectively cured following initial surgical treatment, but it is difficult to predict how an individual patient will actually respond. Recent studies of mammographically screened breast cancers that were stratified into high-, low-, and ultra-low risk of recurrence reported that almost 70% of breast cancers are biologically low risk; and that 15% to 25% of patients may be over treated^{250,251}. It is clear that patients do not respond to treatment in the same way, and many do not benefit from the therapies they are given²⁵² (**Fig 6.2**). Lazarou *et al* (1998) extrapolated from a meta-analysis of adverse drug reactions in prospective studies performed in the USA. In this they reported that over 2 million patients may have had adverse reactions to the drugs they were prescribed²⁵³, and over 100,000 patients may have died. They concluded that adverse drug reactions are the fourth leading cause of death in the US after heart disease, cancer and stroke.

Personalised medicine offers significant opportunities for treating cancer

In chapter 3 I highlighted some of the recent findings from the ICGC pancreatic cancer sequencing project⁶². They reported finding *ERBB2* amplification in 2% of cases and suggested that patients should be recruited to clinical trials for trastuzumab therapy in pancreatic cancer, based on initial screening with IHC as a cost-effective approach. The approach of using the molecular status of tumours to recruit patients with heterogeneous cancer diagnosis into clinical trials is being termed a “basket trial”, and this is likely to become a standard tool as molecular testing becomes more routine.

It is very likely that other ICGC projects will find many cases with *ERBB2* amplification that could potentially be treated with trastuzumab. Using data from the ICGC it is possible to estimate the prevalence of *ERBB2* amplification, and *BRAF* V600E mutation, and *EGFR* mutation or amplification in the major cancers. Combining this information with annual incidence statistics on these same cancers from Cancer Research UK it is possible to estimate the number of patients that could be eligible for the matched targeted therapies. Around 6000 breast cancer patients were eligible for trastuzumab treatment; over 3500 non-breast cancers might be treatable given the prevalence of *ERBB2* amplifications and UK incidence statistics (**Fig 6.3**). Around 5000 melanoma patients were eligible for vemunafirib²⁵⁴ treatment, around 5000 non-melanoma cancers might be treatable given prevalence and incidence statistics (**Fig 6.4**). Around 5000 lung cancer patients were eligible for erlotinib treatment, over 9500 non-

lung cancers might be treatable given prevalence and incidence statistics (**Fig 6.5**). It is clear that many patients could be eligible for targeted therapies (subject to proven clinical utility).

Improved testing is part of the answer

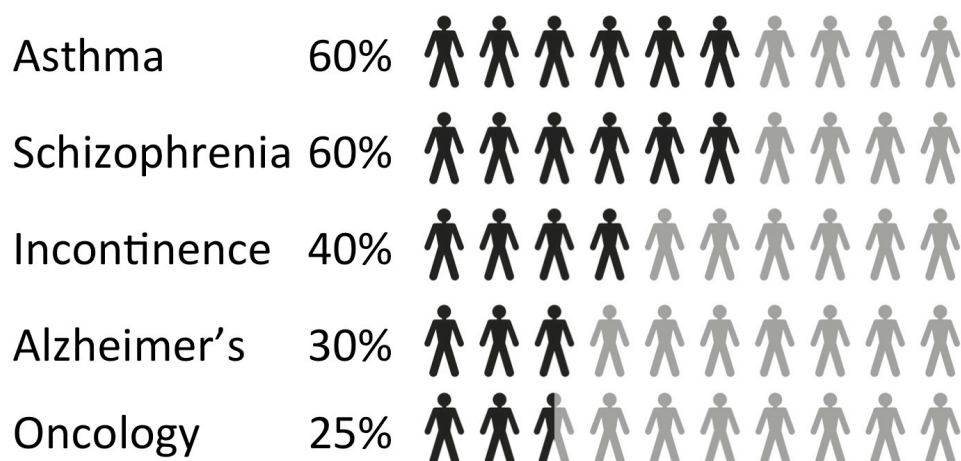
Identifying the patients above for “off-label” use, e.g. non-breast cancer patients for trastuzumab therapy using the current standard of IHC, may be impracticable⁶². Multiple IHC, FISH, PCR or Sanger sequencing tests would need to be run to test all cancer patients for the major driver mutations: these tests could now be combined into a single NGS amplicon panel. Early comparison studies of *KRAS* testing demonstrated 96% concordance between standard methods and NGS in 486 patients. Additionally the NGS results reported mutations not currently tested for that may be clinically relevant²⁵⁵. The sensitivity and specificity of NGS panel testing has also been recently reported suggesting that NGS panel tests for ctDNA may be practicable^{256–258}. Bettegowda *et al* (2013) analysed sensitivity and specificity of mutant *KRAS* detection in colorectal cancers and reported 87% and 99% respectively in ctDNA and reported high concordance between sequencing of tumour tissue and ctDNA²⁵⁶.

There are currently 163 drugs with a gene target that indicates a specific sub-group will need to be identified: 80% are for oncology, psychiatry, infectious diseases, neurology, cardiology & endocrinology and 37% were for cancer treatment²⁵⁹. The list of cancer drivers is only 45 genes long and just five of these account for 50% of the total drug indications. The three commonly prescribed cancer therapies discussed earlier; trastuzumab, vemurafenib and erlotinib, could be combined into a single testing regime of two IHC amplification tests and two mutation screening by sequencing tests, alternatively a single very small NGS amplicon panel test could give the same results, and be significantly faster and cheaper than the three combined tests. A ctDNA based NGS amplicon panel test should be easily deployable in health care systems. As it is minimally invasive tests could be performed on blood drawn in a general practice surgery, with results being ready for an oncologist shortly after. NGS-testing is already being deployed in clinical trials, with over 30 registered at www.clinicaltrials.gov.

Limitations in the molecular analysis of cancer samples: nucleic acid quality

The ideal material to work with for molecular studies is fresh-frozen (FFZN) tumour tissue, as nucleic acids are of high quality. However most cancer samples are preserved in formalin for pathological analysis and stored as formalin-fixed paraffin-embedded (FFPE) blocks, preserving tissue morphology but damaging nucleic acids. The most common artefacts are C>T base substitutions caused by deamination of cytosine bases converting them to uracil and

Fig 6.2: Efficacy rates for different therapeutics



The percentage of patients in different disease populations for which a particular drug class is effective. Molecular targeting of cancer therapeutics is likely to improve their efficacy.

Data from Spear *et al* **Trend Mol Med** 2001 ⁽²⁵²⁾

Fig 6.3: The potential for trastuzumab treatment

Cancer	UK Cases	Percent ERBB2 amplified	Cases ERBB2 amplified
Breast	50285	12.6%	6311
Bowel	41581	3.1%	1289
Lung	43463	2.3%	1000
Bladder	10399	5.8%	607
Prostate	41736	0.9%	384
Pancreas	8773	2.0%	175
Skin	13348	0.6%	80
Brain	9365	0.2%	14
Kidney	10144	0.1%	13
NHL	12783		0

ERBB2 is amplified in 12.5% of Breast cancers as studied by ICGC (USA-TCGA) resulting in 6275 UK cases eligible for trastuzumab therapy. Of the cancers where ICGC report *ERBB2* copy-number amplification status, and where UK incidence statistics are available; a possible 3562 UK cancer patients are potential candidates for trastuzumab therapy. But the majority of these are currently not tested.

Data from www.cbioportal.org.

Fig 6.4: The potential for vemurafenib treatment

Cancer	UK Cases	Percent BRAF V600E	Cases BRAF V600E
Skin	13348	38.8%	5179
Bowel	41581	8.0%	3340
Breast	50285	1.5%	754
Lung	43463	1.4%	608
Brain	9365	1.0%	91
Prostate	41736	0.2%	83
Bladder	10399	0.3%	26
Kidney	10144	0.1%	12
Pancreas	8773	0.0%	0
NHL	12783	-	-

BRAF is mutated in 38.8% of melanomas as studied by ICGC (USA-TCGA) resulting in 5179 UK cases eligible for vemurafenib therapy. Of the cancers where ICGC report *BRAF* V600E mutation status, and where UK incidence statistics are available; a possible 4916 UK cancer patients are potential candidates for vemurafenib therapy. But the majority of these are currently not tested.

Data from www.cbioportal.org.

Fig 6.5: The potential for erlotinib treatment

Cancer	UK Cases	Percent EGFR mut/amp	Cases EGFR mut/amp
Lung	43463	11.8%	5129
Brain	9365	42.5%	3980
Bowel	41581	4.8%	1996
Prostate	41736	2.5%	1043
Breast	50285	1.9%	955
Skin	13348	5.4%	721
Bladder	10399	5.7%	593
Pancreas	8773	1.4%	123
Kidney	10144	1.0%	101
NHL	12783	-	-

EGFR is mutated or amplified in 11.8% of lung cancers as studied by ICGC (USA-TCGA) resulting in 5129 UK cases eligible for erlotinib therapy. Of the cancers where ICGC report *EGFR* mutation and/or amplification status, and where UK incidence statistics are available; a possible 9513 UK cancer patients are potential candidates for erlotinib therapy. But the majority of these are currently not tested.

Data from www.cbioportal.org.

generating thymines during PCR amplification⁷¹, strand-breaks and fragmentation. These reduce the amount of correctly amplifiable template DNA in a sample and must be considered when designing NGS experiments. DNA damage can also occur during the fragmentation of DNA for next-generation sequencing library preparation²⁶⁰. Whilst this damage can be repaired^{261,262}, the use of FFPE DNA in particular is still regarded as difficult. The fragmentation of DNA in FFPE tissue is similar to that seen in ctDNA. For both amplicon- and exome-sequencing this can be mitigated by designing PCR amplicons to be under 150-200bp in length, or by accepting sub-optimal exome library quality control metrics.

Limitations in the molecular analysis of cancer samples: tumour heterogeneity and stromal contamination

Tumour clonal heterogeneity is extensive⁶⁴, and we have shown that tumour evolution during therapy can also be due to heterogeneity in the primary tumour, as revealed by changes in mutant allele frequency during treatment⁹. Microarrays have a limit of detection that will not allow minor clones to be analysed. NGS appears to be limited by the error rates of PCR and sequencing meaning that tumour heterogeneity can currently be assessed by analysis of mutant allele frequency, with a limit of detection of >2%⁸. Understanding the extent of heterogeneity and full tumour burden of a patient at diagnosis is likely to impact their treatment, as this will allow combination therapies directed to multiple molecular targets. Tumour samples are also often heterogeneous with respect to containing normal cells of various types. Since mutant allele frequency, copy-number and gene expression measures all count nucleic acids indiscriminately it is important to identify this and design experiments to be robust to it. Without this then detection sensitivity can be compromised. Samples with high-tumour content (>70%), were selected for inclusion in the METABRIC⁷ breast cancer study.

Understanding cancer biology is vital

The ICGC and other cancer projects using NGS have discovered new insights into cancer biology. However, careful evaluation of the results and clinical trials testing is required to elucidate the impact of these findings. At least one key targeted therapy has failed when applied to a different cancer setting. Vemurafenib increases overall survival rates for 80% of melanoma patients with the *BRAF* V600E mutation²⁴⁶, however many patients quickly become resistant. In colorectal cancers with the *BRAF* V600E mutation the response to Vemurafenib was only 5%, a functional analysis reported that this was due to activation of *EGFR* and recommended clinical trials of combined *BRAF* and *EGFR* inhibitors²⁶³. This

example underlies the importance of understanding the biology of targeted therapeutics, and the likelihood that more complex combined therapies will be required to elicit a long-term response, particularly in heterogeneous disease.

The future for NGS in cancer genomics and companion diagnostics

The development of multiplex-tests that assay hundreds of genes involved in cancer demonstrates the relative ease with which the technology can be applied¹⁴⁷. However there remain significant challenges in deploying these tests in research laboratories. Translating them to general practice is likely to be significantly harder. But the potential for impacting cancer treatment, by identifying patients who could be prescribed treatments, may be so high that it seems likely for multiplex genetic testing by next-generation sequencing to become standard practice in the next decade.

Whilst the advent of a \$1000 genome makes WGS attractive, it is more likely that a tiered approach will be taken and that this might be different for different diseases. WGS is expensive and generates lots of data creating an analysis headache; exomes are cheaper and easier to analyse but may miss variants outside the captured regions; and amplicons are likely to be the fastest and cheapest method but only analyse a small portion of the genome²⁶⁴. It is also possible to sequence amplicons much deeper than exomes, and exomes more deeply than genomes impacting the analysis of mutant allele frequency and tumour heterogeneity. Studies are already being designed to look more carefully at the clinical impact of heterogeneity²⁶⁵.

ctDNA analysis by amplicon-, exome- or whole genome sequencing may also have potential to be used for minimal residual disease (MRD) analysis of solid tumours. MRD is assessed by multiple methods in leukaemia^{266,267}, many of which are unlikely to be applicable to solid tumours, e.g. flow-cytometry. However the concept of using response assessment early in treatment to distinguish prognostic subgroups, by quantifying the change in mutant allele fraction of ctDNA, appears promising. Tumour evolution is also an issue in leukaemia, and this affects current methods based on single markers, the adoption of the amplicon- or exome-sequencing methods described in this thesis may improve the robustness of these MRD tests. The use of a ctDNA MRD test for solid tumours could become a routine tool for longitudinal follow up of patients; if performed with exome-sequencing then the data generated may also reveal new mechanisms of resistance to therapy.

Summary

In this thesis I have described my work on the development of a PCR-based diagnostic test for *ERBB2* amplification in breast cancer. This work preceded the development of trastuzumab therapy but it highlights some of the issues in developing tests that might be used as companion diagnostics. I have presented a review of methodological comparisons, and summarised our two major microarray comparisons. The review highlighted the need to use carefully designed technological comparisons to select platforms for genomic analysis, and the papers described some of the pitfalls of comparison studies. I have also described my contribution to highly cited work developing novel next-generation sequencing technologies for amplicon- and exome-sequencing from tumour and circulating tumour DNA, and described their use in disease monitoring as liquid biopsies. Whilst this work is exciting, the field of circulating tumour DNA analysis and its application to patient treatment and management is still in its infancy. In total this thesis covers almost twenty years of (very enjoyable) work by the author.

Definitions

Abbreviations used commonly in the text are defined here alphabetically and are also given in full at the first instance followed by the abbreviation below in brackets.

aCGH:	Array comparative genomic hybridisation
cfDNA:	Cell-free DNA
ChIP:	Chromatin Immuno-Precipitation
ChIP-seq:	Chromatin Immuno-Precipitation sequencing
CML:	Chronic Myeloid Leukaemia
CNV:	Copy-number variation
CTC:	Circulating tumour cell
ctDNA:	Circulating-tumour DNA
d-PCR:	Differential polymerase chain reaction
ddNTP:	Dideoxy-nucleotide triphosphate
DGE:	Differential gene expression
Exome:	The protein coding portion of the genome
FDA:	United States Food and Drug Administration
FFPE:	Formalin-Fixed Paraffin Embedded
FFZN:	Fresh Frozen
FISH:	Fluorescence in-situ hybridisation
HGP:	Human Genome Project
ICGC:	International Cancer Genome Consortium
IHC:	Immuno-histochemistry
InDel:	Insertion-Deletion
LDT:	Laboratory developed test
LOH:	Loss of heterozygosity
MAQC:	Microarray Quality Control Consortium
MIAME:	Minimum Information About a Microarray Experiment

MINSEQE:	Minimum Information about a high- throughput Nucleotide SeQuencing Experiment
MIQE:	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
miRNA:	Micro RNA
MRD:	Minimal Residual Disease
NGS:	Next-generation sequencing
NICE:	National Institute for Health and Clinical Excellence
NIPT:	Non-invasive prenatal testing
PCR:	Polymerase chain reaction
QA:	Quality Assurance
QC:	Quality Control
qPCR:	Quantitative real-time polymerase chain reaction
RNA-seq:	RNA-sequencing
SNP:	Single nucleotide polymorphism
snpCGH:	Single nucleotide polymorphism comparative genomic hybridisation
SNV:	Single nucleotide variation
TAm-seq:	Tagged Amplicon deep sequencing
TMA:	Tissue microarray
WGS:	Whole genome sequencing

Glossary

Terms used commonly in the text that require further explanation are defined here alphabetically any abbreviations are also defined in full at their first instance in the text followed by the abbreviation in brackets.

Adenocarcinoma: A cancer that develops from glandular epithelium lining tubes, e.g. ductal carcinoma in situ is the most common form of breast cancer.

Adjuvant treatment: A treatment that is given together with, or following, another treatment. Commonly refers to chemotherapy after surgery.

Allele-specific PCR: A type of PCR where one primer is designed to a polymorphic region of the genome. Under stringent PCR conditions only the specific variant will amplify.

Amplification (of gene): An increase in the number of copies per cell of a gene, usually due to whole chromosome aneuploidy, or amplification of a specific region, e.g. ERBB2.

Aneuploidy: The state where chromosome count is not the expected number for a species, e.g. one (or more) extra or missing chromosomes.

Array Comparative Genomic Hybridisation (aCGH): A microarray-based technique for copy-number analysis. Sample and control genomic DNAs are labeled with differently coloured dyes and co-hybridised to a microarray, the relative intensity of each microarray probe is used to infer copy number changes between the sample and the control.

Basket trial: A clinical trial where patients are recruited based on molecular characteristics versus clinical ones, e.g. BRAF V600E status.

Benign: A tumour that does not have the ability to invade or metastasise.

Biomarker: A biological and quantifiable indicator of some biological state or condition that correlates with the presence of particular types and/or sub-types of cancer.

Cancer: A group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body.

Cancer - grade: The description of a cancers appearance.

Cancer - stage: The description of a cancers size, its invasiveness and how far it has spread from where it originated.

Cancer - type: The specific form of cancer a patient has.

Cancer clone: A single molecularly identifiable cancer within a patient. Usually a population of cells derived from a single tumour cell, so that every cell is genetically identical.

Cancer genomics: The study of cancer genomes.

Cancer stem cell: A cell within a cancer which can both self-renew, and give rise to the other cells in the cancer.

Carcinoma: A cancer that develops from epithelial cells.

cDNA: Complementary DNA (cDNA) is DNA synthesized from a mRNA in a reverse transcription and polymerase reaction.

ChIP-seq: Next-generation sequencing of chromatin immunoprecipitated DNA (ChIP): used to identify protein:DNA interactions

Chromatin - Euchromatin: A lightly packed form of chromatin generally associated with active transcription

Chromatin - Heterochromatin: A densely packed form of chromatin generally associated with inactive transcription

Chromatin: A complex of DNA and associated proteins, that is used to package DNA inside the nucleus in chromosomes.

Circulating tumour DNA: Tumour DNA found in blood plasma that is free from the nucleus.

CNV: Copy-number variation (CNV) is a form of structural variation where the DNA is amplified or deleted compared to a reference.

Companion diagnostic test: An in vitro diagnostic test, developed for use by any laboratory, that provides information that is essential for the safe and effective use of its corresponding therapeutic.

Confounding factor: An extraneous variable in a statistical model that correlates with the variable being interrogated, but that is not causal.

Correlation: A statistical technique which tells us if two variables are related, e.g. a specific treatment and cancer growth.

Differential Gene Expression (DGE): An analysis of the variation in quantitative mRNA transcription of two or more biological states reported.

Differential-PCR: A semi-quantitative method to determine DNA amplification status that co-amplifies a target gene with a reference control of known copy-number.

Epigenetics: The study of heritable changes that are not caused by variation in the DNA sequence; e.g. DNA methylation, chromatin modifications, etc.

Exome sequencing: Next-generation sequencing of the exonic portion of the genome. Often performed by capturing exonic regions of a whole genome shotgun library for NGS by in-solution hybridisation to biotinylated "exon baits".

Exome: The sum total of all the exons in the genome, often including regulatory sequences as well.

Exon: A nucleotide sequence encoded by a gene that remains present within the final mature RNA product of that gene after introns have been removed by RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts. Includes but is not limited to protein-coding regions.

FFPE: Formalin fixed paraffin embedded: a tissue preservation technique commonly used in pathology laboratories that results in degradation of nucleic acids.

FFZN: Fresh frozen: a tissue preservation technique which uses rapid freezing in liquid nitrogen to preserve nucleic acid integrity.

FISH: Fluorescence in-situ hybridisation: a technique used to locate specific DNA (and RNA) sequences within a cell.

Gene expression - absolute: The quantitative level at which the DNA from a gene is being transcribed into mRNA.

Gene expression - differential (DGE): An analysis of the variation in quantitative mRNA

transcription of two or more biological states reported

Gene expression signature: A defined set of genes and their expression levels that describes a particular biological state, or tumour sub-class.

Gene: The molecular unit of heredity, usually used to describe a unit of DNA that codes for mRNA and a functional protein.

Genome sequencing: DNA sequencing of the entire genome.

Genome: The complete DNA complement of an individual.

Genomics: The study of genomes.

Gold standard: The test that is the best available under the conditions being considered.

Human genome reference sequence: The genome sequence used as a reference point for genomic studies, currently the Genome Reference Consortium human genome (build 37) which is a haploid mosaic of DNA sequences from 37 donors.

IHC: Immunohistochemistry: a laboratory technique used to detect the presence of specific protein or markers on cancer tumours.

International Cancer Genome Consortium (ICGC): A collaboration of cancer researchers launched in 2008 to coordinate large-scale next-generation sequencing based cancer studies in tumours from 50 cancer types and/or subtypes that are of main importance across the globe.

Laboratory developed test (LDT): An in vitro diagnostic test developed for use in a single laboratory, but one that is not regulated in the same way as a companion diagnostic test.

Leukaemia: A cancer that develops from the blood.

Loss of heterozygosity (LOH): A gross chromosomal event that results in loss of an entire gene or chromosomal region in one of the parental alleles, resulting in the presence of single alleles (AA or BB) rather than the usual two (AA, AB, BB across that region).

Lymphoma: A cancer that develops from the lymph nodes, there are two main types: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL)

Malignancy/metastasis: A state describing cancer which has the capability to spread, or has

actually spread from the primary site to secondary sites

Melanoma: A cancer that develops from the pigment-containing cells in the skin (melanocytes).

Methylome: The complement of all nucleic acid methylation modifications of the DNA in an organism's genome, generally in a specific sample or cell-type from an individual.

Micro-RNA: A single-stranded, non-coding form of RNA, having only about 20-30 nucleotides, that has a number of functions including the regulation of gene expression.

Microarray: A method that allows thousands of DNA or RNA sequences to be assessed in a single experiment, e.g. DGE or CNV.

Minimal residual disease: The term used to describe the small numbers of leukaemic cells that remain in the patient after treatment.

Molecular pathology: The study and diagnosis of disease through the examination of nucleic acids within cells, tissues, organs or bodily fluids; often using molecular techniques such as PCR, NGS or microarrays.

Mutant allele frequency: The proportion of a particular mutant allele (non-normal variant of a gene) in a tumour, or in a blood sample from a suspected cancer patient.

Mutation: A change in the DNA sequence compared to a reference, usually associated with disease and distinct from normal variation (polymorphism).

Mutation - driver: A mutation present in cancer that is responsible for tumorigenesis on its own, or in combination with a small number of other driver mutations. A mutation that confers a selective growth advantage on the cell and thus is required for cancer to develop.

Mutation - germline: A mutation present in the gametes, or zygote a very early stage of development, and presumed to be present in every cell of an individual.

Mutation - missense: A mutation that alters the amino acid sequence of the encoded protein, often rendering it non-functional.

Mutation - non-synonymous: A mutation that alters the amino acid sequence of a protein.

Mutation - nonsense: A mutation that causes a premature stop codon in a sequence resulting

in truncation of translation and non-functional protein.

Mutation - passenger: A mutation present in cancer that confers no selective growth advantage on the cell and thus does not contribute to cancer. Often present as a result of increased mutagenesis in cancer genomes.

Mutation - somatic: A mutation present in a cancer that is not present in the normal cells of an individual; i.e. white blood cell DNA in non myeloid malignancies.

Mutation hot-spot: A region of DNA where mutations accumulate more than would be expected by chance.

Next-generation sequencing: DNA sequencing methods that can be used to sequence the whole genome, transcriptome or methylome of an individual, distinct from the Sanger sequencing used to complete the HGP.

Oncogene: A gene that has the potential to cause cancer, e.g. KRAS. Mutation of oncogenes often leads to their activation promoting uncontrolled cell proliferation.

Orthogonal method: An independent method used to validate findings from an experiment, e.g. qPCR for RNA-seq.

Penetrance: The percentage of patients with a disease, e.g. cancer sub-type, that carry a specific variant of a gene. A mutation with 95% penetrance will cause 95% of individual with that mutation to develop the disease, whilst 5% will not.

Personalised medicine: Treatment directed towards an individual patients requirements, often based on clinico-pathological or molecular phenotypes. Contrasts with stratified medicine, which directs treatment to groups of patients.

Philadelphia chromosome : A specific chromosomal abnormality associated with chronic myelogenous leukemia (CML). It is the result of a reciprocal translocation between the BCR and ABL genes creating a constitutively activated tyrosine kinase.

Polymerase Chain Reaction: A method used to amplify a single copy or a few copies of a DNA sequence (or RNA through reverse transcription).

Polymerase Chain Reaction - qPCR: A method used to simultaneously amplify and quantify a DNA sequence (or RNA through reverse transcription) in a sample.

Primary cancer: The site where a cancer is suspected of originating.

Prognostic/predictive factors: A prognostic factor is one that is objectively measurable and provides information on the likely outcome of disease in an individual, prognostic factors define the effects of patient or tumor characteristics on the patient outcome. A predictive factor is one that provides information on the likely benefit from treatment, predictive factors define the effect of treatment on the tumor.

Proto-oncogene: The normal cellular version of a gene which, when mutated, can become an oncogene.

Pseudogene: A copy of a gene that has been mutated into an inactive form through evolution, but that can confound DNA sequence based techniques in the laboratory, e.g. by affecting PCR, or in analysis, e.g. by affecting alignment to a reference genome.

Randomisation: The random allocating of experimental samples across groups, e.g. treatment versus control. It is often used to reduce the impact of confounding factors in a formal experimental design.

Reflex-testing: Follow-up testing automatically initiated when certain test results are observed in the laboratory; used to clarify or elaborate on primary test results

Replicate: A single measurement from an experimental condition, replication allows the biological variability to be estimated

Replication - biological: A replicate from an independent biological samples, i.e. different patients with the same disease.

Replication - technical: A replicate from a non-independent biological sample, i.e. different blood-draws from the same patient.

RNA-seq: Next-generation sequencing of RNA: used to identify DGE or alternative splicing of mRNA

Sanger sequencing: The method of DNA sequencing invented by Frederick Sanger, used to complete the HGP.

Sarcoma: A cancer that develops from the mesenchyme (cells of the connective tissue).

Secondary cancer: A tumour that has spread from a primary site to a different site or organ,

but one that will still be referred to by its likely site of primary origin.

Sensitivity & Specificity - Sensitivity: The proportion of samples known to be positive for a test, which actually test positive.

Sensitivity & Specificity - Specificity: The proportion of samples known to be negative for a test, which actually test negative.

Sequencing coverage/depth: The number of times a nucleotide is read during the sequencing process, often used interchangeably.

Single nucleotide polymorphism/variant (SNP/SNV): A single base pair change in the DNA sequence detected from the reference genome. A variant may be benign (a polymorphism) or pathogenic (a mutation).

SNP-CGH: A microarray based technique for copy-number analysis. Sample and control genomic DNAs are genotyped using microarrays, the relative intensities microarray probes is used to infer copy number changes between the sample and the control. The genotype calls can also be used to infer LOH making snpCGH a more powerful technique than aCGH.

Stratified medicine: Treatment directed towards groups of patients, often based on clinico-pathological or molecular phenotypes. Contrasts with personalised medicine, which directs treatment to individual patients.

Tam-seq: A method for sequencing PCR amplicons using next-generation sequencing

Targeted therapy: A drug or treatment that specifically targets cancer cells based on their molecular subtype, patients are usually selected by the use of a companion diagnostic test e.g. ERBB2 amplification status and Herceptin.

The Cancer Genome Atlas: A project, launched by the NIH in 2005, to catalogue genetic mutations responsible for cancer (now part of the ICGC)

Transcriptome: The complement of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA of an individual, generally in a specific sample or cell-type from an individual.

Trio analysis: A specialised genome analysis that compares results from both parent to an affected child with the aim of identifying the genetic features that may be causal of a disease;

often completed using exome sequencing.

Tumour heterogeneity: The observation that different tumour cells can show distinct genotypic and phenotypic profiles. This can occur between tumours (inter-tumour heterogeneity) and within tumours (intra-tumour heterogeneity).

Tumour suppressor gene: A gene that protects a cell from tumourigenesis, e.g. TP53. Mutation of tumour suppressors often leads to their inactivation allowing uncontrolled cell proliferation.

Tumour/clonal evolution: The process of change in a tumour driven by the accumulation of mutations whereby certain clones can become dominant over others, this process can be driven by treatment and is one model for the development of cancer.

Wild-type: The genotype of reference allele of a species, more often the most common genotype in the population.

References

1. Jennings, B. A., Hadfield, J., Worsley, S. D., Girling, A. & Willis, G. A differential PCR assay for the detection of c-erbB 2 amplification used in a prospective study of breast cancer. *Mol. Pathol.* **50**, 254–256 (1997).
2. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**, 240–8 (2009).
3. Curtis, C. *et al.* The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* **10**, 588–611 (2009).
4. Git, A. *et al.* Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16**, 991–1006 (2010).
5. Lynch, A. *et al.* The cost of reducing starting RNA quantity for Illumina BeadArrays : A bead-level dilution experiment . *BMC Genomics* **11**, 540–9 (2010).
6. Aldridge, S. & Hadfield, J. Introduction to miRNA Profiling Technologies and cross-platform comparison. *Methods Mol Biol* **822**, 19–31 (2012).
7. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–52 (2012).
8. Forsheew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012).
9. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–12 (2013).
10. Idris, S. F., Ahmad, S. S., Scott, M., Vassiliou, G. & Hadfield, J. The role of high-throughput technologies in clinical cancer genomics. *Expert Rev. Mol. Diagn.* **13**, 167–81 (2013).
11. Azizan, E. A. B. *et al.* Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension. *Nat. Genet.* (2013). doi:10.1038/ng.2716
12. Hadfield, J. & Eldridge, M. D. Multi-genome alignment for quality control and contamination screening of next-generation sequencing data. *Front. Genet.* **20**, 31 (2014).
13. Sanger, F. The Free Amino Groups of Insulin. **39**, 507–15 (1945).
14. Edman, P. Method for the determination of the amino acid sequence in peptides. *Acta Chem. Scand.* 283–93 (1950).
15. Sanger, F. & Tuppy, H. The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 1. *Biochem. J* **49**, 463–81 (1951).

16. Sanger, F. & Tuppy, H. The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 2. *Biochem. J* **49**, 481–90 (1949).
17. Brownlee, G. G., Sanger, F. & Barrell, B. G. The Sequence of 5 s Ribosomal Ribonucleic Acid. *J. Mol. Biol* **34**, 379–412 (1968).
18. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–8 (1975).
19. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–95 (1977).
20. Sanger, F. & Nicklen, S. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
21. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* 2399–2412 (1986).
22. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* (80-.). **238**, 336–41 (1987).
23. Smith, L. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–9 (1986).
24. Rosenblum, B. B. *et al.* New dye-labeled terminators for improved DNA sequencing patterns. **25**, 4500–4504 (1997).
25. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Genetics* **5**, 335–44 (2004).
26. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
27. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
28. Rothberg, J. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
29. Bentley, D. R., Balasubramanian, S., Swerdlow, H., Smith, G. & Milton, J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
30. Retief, J. (Illumina). *For all you seq.* 1 (2014).
31. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
32. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial

- genome. *Science* (80-.). **309**, 1728–1732 (2005).
33. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* (80-.). **323**, 133–138 (2008).
 34. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* (80-.). 1–7 (2009).
 35. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–60 (2001).
 36. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039–43 (2002).
 37. Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
 38. Gencode_Consortium. Gencode v20 (April 2014 freeze, GRCh38). (2014). at <<http://www.gencodegenes.org/stats.html>>
 39. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–74 (2012).
 40. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
 41. Sakharkar, M. K., Chow, V. T. K. & Kanguane, P. Distributions of Exons and Introns in the Human Genome. **4**, 387–393 (2004).
 42. Zhu, L. *et al.* Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**, 47 (2009).
 43. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 19–21 (2008). doi:10.1038/ng.259
 44. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–54 (2005).
 45. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14**, 880–93 (2013).
 46. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, (2001).
 47. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–51 (2001).
 48. Waterston, R. H., Lander, E. S. & Sulston, J. E. More on the sequencing of the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3022–4; author reply 3025–6 (2003).

49. Bernstein, C., Prasad, A. R., Nfonsam, V. & Bernstein, H. DNA Damage , DNA Repair and Cancer. (2013).
50. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
51. Fearon, E. F. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
52. Knudson, a G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* **1**, 157–62 (2001).
53. Wilentz, R. E. *et al.* Loss of Expression of Dpc4 in Pancreatic Intraepithelial Neoplasia : Evidence That DPC4 Inactivation Occurs Late in Neoplastic Progression. *Cancer Res.* **60**, 2002–2006 (2000).
54. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer : The Next Generation. *Cell* **144**, 646–674 (2011).
55. Cancer-Research-UK. Cancer statistics: Key Facts. (2014).
56. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* (2009). doi:10.1038/nature08658
57. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
58. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–8 (2010).
59. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–93 (2012).
60. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
61. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Rep.* **7**, 1740–52 (2014).
62. Chou, A. *et al.* Clinical and molecular characterization of HER2 amplified-pancreatic cancer. *Genome Med.* **5**, 78 (2013).
63. Weaver, J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–43 (2014).
64. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366**, 883–92 (2012).
65. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–74 (2006).
66. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational

signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).

67. Knudson, A. G. Mutation and Cancer : Statistical Study of Retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
68. Fisher, R. The arrangement of field experiments. *J Minist Agric Gt. Britain* **33**, 503–513 (1926).
69. Dako. HercepTest Interpretation Manual - Breast. (2002).
70. Kearney, H. M. *et al.* American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities. *Genet. Med.* **13**, 676–9 (2011).
71. Do, H. & Dobrovic, A. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. *Oncotarget* **3**, 546–558 (2012).
72. Srinivasan, M. & Sedmak, D. Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids. **161**, 1961–1971 (2002).
73. Higuchi, R., Dollinger, G., P, W. S. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Nat. Biotechnol.* **10**, 413–417 (1992).
74. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Nat. Biotechnol.* **11**, 1026–1030 (1993).
75. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
76. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. (1986).
77. Bustin, S. A. *et al.* The MIQE Guidelines : Minimum Information for Publication of Quantitative Real-Time PCR Experiments SUMMARY : *Clin. Chem.* **622**, 611–622 (2009).
78. Rutledge, R. G. & Cote, C. Mathematics of quantitative kinetic PCR and the application of standard curves. *Nucleic Acids Res.* **31**, (2003).
79. Garner, D., Johnson, L., Yue, S., Roth, B. & Haugland, R. Dual DNA staining assessment of bovine sperm viability using SYBR-14 and propidium iodide. *J. Androl.* **15**, 620–629 (1994).
80. Schneeberger, C., Speiser, P., Kury, F. & Zeillinger, R. Quantitative detection of reverse transcriptase-PCR products by means of a novel and sensitive DNA stain. *PCR Methods Appl.* **4**, 234–8 (1995).
81. Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5'-3' exonuclease activity of

- Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 7276–80 (1991).
82. Gibson, U. E., Heid, C. a & Williams, P. M. A novel method for real time quantitative RT-PCR. *Genome Res.* **6**, 995–1001 (1996).
 83. Bieche, I. *et al.* Real-Time Reverse Transcription-PCR Assay for Future Management of ERBB2 -based Clinical Applications. *Clin. Chem.* **45**, 1148–1156 (1999).
 84. Beyser, K., Reiser, A., Gross, C., Tabiti, K. & Gmbh, R. D. Real-time Quantification of HER2/neu Gene Amplification by LightCycler Polymerase Chain Reaction (PCR) – a New Research Tool. *Biochem.* **922**, 15–18 (2001).
 85. Pawlowski, V., Revillion, F., Hornez, L. & Peyrat, J. A real-time one- step reverse transcriptase-polymerase chain reaction method to quantify c-erbB-2 expression in human breast cancer. *Cancer Detect. Prev.* **24**, 212–223 (2000).
 86. O'Malley, F. P. *et al.* Comparison of HER2/neu status assessed by quantitative polymerase chain reaction and immunohistochemistry. *Am. J. Clin. Pathol.* **115**, 504–11 (2001).
 87. Bernard, P. S. & Wittwer, C. T. Real-time PCR technology for cancer diagnostics. *Clin. Chem.* **48**, 1178–85 (2002).
 88. Southern, E. Detection of specific sequences among DNA fragments separated by gel electrophoresis. 1975. *J. Mol. Biol* **98**, 503–517 (1975).
 89. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *PNAS* **74**, 5350–4 (1977).
 90. Towbin, H., Staehelint, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets : Procedure and some applications. *PNAS* **76**, 4350–4354 (1979).
 91. Schena, M., Shalon, D., Davis, R. W. & Brown, P. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science (80-.)*. **270**, 467–470 (1995).
 92. Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *PNAS* **91**, 5022–6 (1994).
 93. Lockhart, D. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotech.* **14**, 1675–1680 (1996).
 94. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
 95. Gunderson, K. L. Decoding Randomly Ordered DNA Arrays. *Genome Res.* **14**, 870–877 (2004).

96. Yeakley, J. M. *et al.* Profiling alternative splicing on fiber-optic arrays. **20**, 353–358 (2002).
97. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–8 (2008).
98. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).
99. Fan, J.-B. Next-Generation MicroRNA Expression Profiling Technology. *Methods Mol. Biol.* **XI**, (2012).
100. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (80-.).* **258**, 818–21 (1992).
101. Mei, R. Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. *Genome Res.* **10**, 1126–1137 (2000).
102. Lindblad-toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18**, 1001–1005 (2000).
103. Dumur, C. I. *et al.* Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. *Genomics* **81**, 260–269 (2003).
104. Golub, T. R. *et al.* Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. *Science (80-.).* **286**, 531–537 (1999).
105. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
106. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
107. Van de Vijver, M. J. *et al.* A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
108. Mardis, E. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
109. Shendure, J., Porreca, G. J. & Church, G. M. Overview of DNA Sequencing Strategies. *Curr. Protoc. Mol. Biol.* **81**, 1–11 (2008).
110. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* (2008). doi:10.1038/nrg2484
111. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific mRNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
112. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–9 (2013).

113. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* (2010). doi:10.1038/nature08903
114. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).
115. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* (80-.). **316**, 1497–1502 (2007).
116. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein – DNA association. *DNA Seq.* **28**, 327–334 (2001).
117. Simon, R. & Roychowdhury, S. Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* **12**, 358–69 (2013).
118. Simon, R. M. & Dobbin, K. Experimental design of DNA microarray experiments. *Biotechniques Suppl.* 16–21 (2003).
119. Wei, C., Li, J. & Bumgarner, R. E. Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics* **5**, 87 (2004).
120. Strand, C., Enell, J., Hedenfalk, I. & Fernö, M. RNA quality in frozen breast cancer samples and the influence on gene expression analysis – a comparison of three evaluation methods using microcapillary electrophoresis traces. *BMC Mol. Biol.* **9**, 1–9 (2007).
121. Shi, L. & MAQC. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
122. Andrews, S. FastQC. (2010). at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
123. Buchdunger, E. *et al.* Inhibition of the Abl Protein-Tyrosine Kinase in Vitro and in Vivo by a 2-Phenylaminopyrimidine Derivative Inhibition of the Abi Protein-Tyrosine Kinase in Vitro and in Vivo by a Derivative. 100–104 (1996).
124. Hochhaus, A. *et al.* Favorable long-term follow-up results over 6 years for response , survival , and safety with imatinib mesylate therapy in chronic-phase chronic myeloid leukemia after failure of interferon- α treatment. **111**, 1039–1043 (2008).
125. Zhen, C. & Wang, Y. L. Molecular monitoring of chronic myeloid leukemia: international standardization of BCR-ABL1 quantitation. *J. Mol. Diagn.* **15**, 556–64 (2013).
126. Coussens, L. *et al.* Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science* (80-.). **230**, 1132–9 (1985).
127. Slamon, D. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the Her2/neu oncogene. *Science* (80-.). **9**, 177–82 (1987).

128. Slamon, D. *et al.* Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* (80-.). **244**, 707–12 (1989).
129. Baak, J. *et al.* Comparative long-term prognostic value of quantitative HER-2/neu protein expression, DNA ploidy, and morphometric and clinical features in paraffin-embedded invasive breast cancer. *Lab Invest* 215–23 (1991).
130. Cobleigh, M. *et al.* Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *J Clin Oncol* **17**, 2639–48 (1999).
131. Bang, Y.-J. *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* **376**, 687–97 (2010).
132. Moelans, C. B., de Weger, R. a, Van der Wall, E. & van Diest, P. J. Current technologies for HER2 testing in breast cancer. *Crit. Rev. Oncol. Hematol.* **80**, 380–92 (2011).
133. Rhodes, A., Jasani, B., Anderson, E., Dodson, A. R. & Balaton, A. J. Evaluation of HER-2/neu immunohistochemical assay sensitivity and scoring on formalin-fixed and paraffin-processed cell lines and breast tumors: a comparative study involving results from laboratories in 21 countries. *Am. J. Clin. Pathol.* **118**, 408–17 (2002).
134. Dekker, T. J. a *et al.* Determining sensitivity and specificity of HER2 testing in breast cancer using a tissue micro-array approach. *Breast Cancer Res.* **14**, 1–12 (2012).
135. Mass, R. D. *et al.* Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab. *Clin. Breast Cancer* **6**, 240–6 (2005).
136. Margolin, A. a *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1 (2013).
137. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* **5**, 181ra50 (2013).
138. Frye RA, Benz CC, L. E. Detection of amplified oncogene by differential polymerase chain reaction. *Oncogene* **4**, 1153–7 (1989).
139. Hubbard, a L., Doris, C. P., Thompson, a M., Chetty, U. & Anderson, T. J. Critical determination of the frequency of c-erbB-2 amplification in breast cancer. *Br. J. Cancer* **70**, 434–9 (1994).
140. An, H. *et al.* ERBB2 gene amplification detected by fluorescent differential polymerase chain reaction in paraffin-embedded breast carcinoma tissues. *Int J cancer* **64**, 291–7 (1995).

141. Johnson, R., Ricci, A. J., Cartun, R., Ackroyd, R. & Tsongalis, G. p185HER2 overexpression in human breast cancer using molecular and immunohistochemical methods. *Cancer Invest* **18**, 336–42 (2000).
142. Bartlett, J., Mallon, E. & Cooke, T. The clinical evaluation of HER-2 status: which test to use? *J. Pathol.* **199**, 411–7 (2003).
143. Tsongalis, G. & Reid, A. J. HER2: The Neu Prognostic Marker for Breast Cancer. *Crit. Rev. Clin. Lab. Sci.* **38**, 167–82 (2001).
144. Naidu, R., Abdul, N. W., Yadav, M., Kutty, M. K. & Nair, S. Detection of amplified int-2/FGF-3 gene in primary breast carcinomas using differential polymerase chain reaction. *Int J Mol Med.* **8**, 193–8 (2001).
145. Naidu, R., Wahab, N., Yadav, M. & Kutty, M. Protein expression and molecular analysis of c-myc gene in primary breast carcinomas using immunohistochemistry and differential polymerase chain reaction. *Int J Mol Med.* **9**, 189–96 (2002).
146. Naidu, R., Wahab, N., Yadav, M. & Kutty, M. Expression and amplification of cyclin D1 in primary breast carcinomas: relationship with histopathological types and clinico-pathological parameters. *Oncol Rep.* **9**, 409–16 (2002).
147. Frampton, G. M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–31 (2013).
148. Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *PNAS* **105**, 16224–29 (2008).
149. Leary, R. J. *et al.* Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. *Sci. Transl. Med.* **2**, (2010).
150. Dawson, S.-J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–209 (2013).
151. FDA. Companion Diagnostics. at <<http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm301431.htm>>
152. NICE. Cetuximab for the first-line treatment of metastatic colorectal cancer. (2012). at <<http://www.nice.org.uk/guidance/TA176>>
153. NICE. KRAS mutation testing of tumours in adults with metastatic colorectal cancer. at <<http://guidance.nice.org.uk/DT/14>>
154. Tejpar, S. *et al.* Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist* **15**, 390–404 (2010).
155. Van Cutsem, E. *et al.* Cetuximab and Chemotherapy as Initial Treatment for

Metastatic Colorectal Cancer. *N. Engl. J. Med.* **360**, 1408–17 (2009).

156. NICE. NICE guidance on vemurafenib for treating locally advanced or metastatic BRAF V600 mutation-positive malignant melanoma. (2012). at <<http://www.nice.org.uk/guidance/TA269>>
157. Personalized Medicine Coalition. The case for personalized medicine. at <http://www.personalizedmedicinebulletin.com/wp-content/uploads/sites/205/2011/11/Case_for_PM_3rd_edition1.pdf>
158. Dvinge, H. *et al.* The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378–82 (2013).
159. Editorial. All things being equal. *Nat. Methods* **9**, 111–111 (2012).
160. Hanneman, S. K. & Katz, J. B. Design, Analysis and Interpretation of Method-Comparison Studies. *AACN Adv Crit Care* **19**, 223–34 (2008).
161. Hartnack, S. Issues and pitfalls in method comparison studies. *Vet. Anaesth. Analg.* **41**, 227–232 (2014).
162. Armbruster, D. & Miller, R. R. The Joint Committee for Traceability in Laboratory Medicine (JCTLM): A Global Approach to Promote the Standardisation of Clinical Laboratory Test Results. *Clin Biochem Rev* **28**, 105–113 (2007).
163. NICE. *Guide to the multiple technology appraisal process.* (2009). at <<http://www.nice.org.uk/About/What-we-do/Our-Programmes/NICE-guidance/NICE-technology-appraisal-guidance>>
164. NICE. *EGFR-TK mutation testing in adults with locally advanced or metastatic non-small-cell lung cancer.* (2013). at <<http://www.nice.org.uk/guidance/DG9>>
165. Mansfield, E., Leary, T. J. O. & Gutman, S. I. Food and Drug Administration Regulation of in Vitro Diagnostic Devices. *J. Mol. diagnostics* **7**, 2–7 (2005).
166. Baker, S. *et al.* External RNA Controls Consortium. *Nat. Methods* **2**, 731–734 (2005).
167. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–51 (2014).
168. Kothapalli, R., Yoder, S. J., Mane, S. & Loughran, T. P. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
169. Kuo, W. P., Jenssen, T., Butte, A. J., Ohno-machado, L. & Kohane, I. S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
170. Piper, M. D. W. *et al.* Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **277**, 37001–8 (2002).

171. Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J. & Sealfon, S. C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, 1–9 (2002).
172. Barczak, A. *et al.* Spotted Long Oligonucleotide Arrays for Human Gene Expression Analysis. *Genome Res.* **13**, 1775–1785 (2003).
173. Hyung-Lae, K. Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34 cells. *Exp. Mol. Med.* **35**, 460–466 (2003).
174. Rogojina, A. T., Orr, W. E., Song, B. K. & Geisert, E. E. Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol. Vis.* 482–496 (2003).
175. Jurata, L. W. *et al.* Comparison of microarray-based mRNA profiling technologies for identification of psychiatric disease and drug signatures. *J. Neurosci. Methods* **138**, 173–88 (2004).
176. Mah, N. *et al.* A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics* **16**, 361–70 (2004).
177. Parrish, M. L. *et al.* A microarray platform comparison for neuroscience applications. *J. Neurosci. Methods* **132**, 57–68 (2004).
178. Shippy, R. *et al.* Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61 (2004).
179. Woo, Y. *et al.* A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms. *J. Biomol. Tech.* **15**, 276–284 (2004).
180. Yauk, C., Berndt, M. L., Williams, A. & Douglas, G. R. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32**, e124 (2004).
181. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
182. Pylatuik, J. D. & Fobert, P. R. Comparison of transcript profiling on Arabidopsis microarray platform technologies. *Plant Mol. Biol.* **58**, 609–24 (2005).
183. Schlingemann, J. *et al.* Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab. Invest.* **85**, 1024–39 (2005).
184. De Reyniès, A. *et al.* Comparison of the latest commercial short and long oligonucleotide microarray technologies. *BMC Genomics* **7**, 51 (2006).
185. Kuo, W. P. *et al.* A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* **24**, 832–840 (2006).

186. Patterson, T. A. *et al.* Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol.* **24**, 1140–1150 (2006).
187. Severgnini, M. *et al.* Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. *Anal. Biochem.* **353**, 43–56 (2006).
188. Chen, J. *et al.* A comparison of microarray and MPSS technology platforms for expression analysis of Arabidopsis. *BMC Genomics* **8**, 414 (2007).
189. Maouche, S. *et al.* Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells. *BMC Genomics* **13**, 1–13 (2008).
190. Hockley, S. L. *et al.* Interlaboratory and interplatform comparison of microarray gene expression analysis of HepG2 cells exposed to benzo(a)pyrene. *OMICS* **13**, 115–25 (2009).
191. Witzel, I. D. *et al.* Comparison of microarray-based RNA expression with ELISA-based protein determination of HER2, uPA and PAI-1 in tumour tissue of patients with breast cancer and relation to outcome. *J. Cancer Res. Clin. Oncol.* **136**, 1709–18 (2010).
192. Su, Z. *et al.* Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem. Res. Toxicol.* **24**, 1486–93 (2011).
193. Xu, X. *et al.* Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC Bioinformatics* **14 Suppl 9**, S1 (2013).
194. Suárez-Fariñas, M. & Magnasco, M. O. Comparing microarray studies. *Methods Mol. Biol.* **377**, 139–52 (2007).
195. Liu, F., Kuo, W. P., Jenssen, T. & Hovig, E. Next Generation Microarray Bioinformatics. *Methods Mol. Biol.* **802**, 141–155 (2012).
196. Ambion. *FirstChoice Human Brain Reference Total RNA*. **6050**, 2 (2005).
197. Stratagene. *Universal Human Reference RNA Cat#740000*, Stratagene, USA. 1 (2005).
198. HapMap-Consortium. The International HapMap project. *Nihon Rinsho*. **63 Suppl 1**, 29–34 (2005).
199. Forozan, F. *et al.* Comparative Genomic Hybridization Analysis of 38 Breast Cancer Cell Lines: A Basis for Interpreting Complementary DNA Microarray Data. *Cancer Res.* **60**, 4519–25 (2000).
200. Peiffer, D. A. *et al.* High-resolution genomic profiling of chromosomal aberrations

using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–48 (2006).

201. Affymetrix. Genome-Wide Human SNP Array 6.0 Sample Data Set. at <http://www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx>
202. Git, A. *et al.* PMC42, a breast progenitor cancer cell line, has normal-like mRNA and microRNA transcriptomes. *Breast Cancer Res.* **10**, 1–16 (2008).
203. Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38**, 1375–1377 (2006).
204. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* **29**, 365–71 (2001).
205. FGED. *MINSEQE: Minimum Information about a high-throughput Nucleotide Sequencing Experiment - a proposal for standards in functional genomic data reporting.* **0**, 1–2 (2012).
206. Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–18 (2013).
207. Xu, H., Eirew, P., Mullaly, S. C. & Aparicio, S. The Omics of Triple-Negative Breast Cancers. *Clin. Chem.* **133**, 122–133 (2014).
208. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144**, 27–40 (2011).
209. Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**, e1003535 (2014).
210. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
211. ICGC data portal. at <<https://dcc.icgc.org>>
212. Almendro, V. *et al.* Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res.* **74**, 1338–48 (2014).
213. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–10 (2012).
214. Lander, E. S. & Waterman, S. Genomic Mapping by Fingerprinting Random Clones : A Mathematical Analysis. *Genomics* **239**, 231–239 (1988).
215. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–5 (2010).
216. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–3 (2010).

217. Gahl, W. & Tifft, C. The NIH Undiagnosed Diseases Program. *JAMA* **305**, 1904–1905 (2011).
218. Rabbani, B., Mahdih, N., Hosomichi, K., Nakaoka, H. & Inoue, I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J. Hum. Genet.* **57**, 621–32 (2012).
219. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–11 (2013).
220. NIH Mendelian Exome Project. at <http://www.nhlbi.nih.gov/resources/genetic/genomics/programs/mendelian.htm>
221. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* **20**, 490–7 (2012).
222. Leon, S. A., Shapiro, B., Sklaroff, D. M. & Yaros, M. J. Free DNA in the Serum of Cancer Patients and the Effect of Therapy. *Cancer Res.* **37**, 646–50 (1977).
223. Anker, P., Mulcahy, H. & Stroun, M. Circulating nucleic acids in plasma and serum as a noninvasive investigation for cancer: time for large-scale clinical studies? *Int. J. Cancer* **103**, 149–52 (2003).
224. Lo, Y. M. D. *et al.* Presence of fetal DNA in maternal plasma and serum. *Lancet* **350**, 485–487 (1997).
225. Lo, Y. M. *et al.* Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* **62**, 768–75 (1998).
226. Lo, Y. M. D. *et al.* Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13116–21 (2007).
227. Fan, H. C. & Quake, S. R. Detection of aneuploidy with digital polymerase chain reaction. *Anal. Chem.* **79**, 7576–9 (2007).
228. Bianchi, D. W. & Wilkins-Haug, L. Integration of noninvasive DNA testing for aneuploidy into prenatal care: what has happened since the rubber met the road? *Clin. Chem.* **60**, 78–87 (2014).
229. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 16368–73 (2005).
230. Jahr, S. *et al.* DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells. *Cancer Res.* 1659–1665 (2001).
231. Chan, K. C. A. *et al.* Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing. *Clin. Chem.* **59**, 211–24 (2013).

232. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Supplementary material. *Nature* **497**, 108–12 (2013).
233. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. Supplementary material. *Sci. Transl. Med.* **6**, 224ra24 (2014).
234. Ahmed, A. A. *et al.* Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *J. Pathol.* **221**, 49–56 (2010).
235. Forbes, S. a *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–50 (2011).
236. Diehl, F. *et al.* BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nat. Methods* **3**, 551–559 (2006).
237. Isakoff, S. J. *et al.* Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. *Cancer Res.* **65**, 10992–11000 (2005).
238. Knudsen, E. S. & Knudsen, K. E. Tailoring to RB: tumour suppressor status and therapeutic response. *Nat. Rev. Cancer* **8**, 714–724 (2008).
239. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2**, e73 (2005).
240. D’Arcangelo, M. & Hirsch, F. R. Clinical and comparative utility of afatinib in non-small cell lung cancer. *Biologics* **8**, 183–92 (2014).
241. Nagalingam, A. *et al.* Med1 plays a critical role in the development of tamoxifen resistance. *Carcinogenesis* **33**, 918–30 (2012).
242. Liu, L. *et al.* Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.* **69**, 6871–8 (2009).
243. NEQAS. NEQAS quality assurance schemes. (2012). at <<http://www.ukneqas-molgen.org.uk/molecular-pathology>>
244. NEQAS. NEQAS quality assurance for next-generation sequencig. (2012).
245. Slamon, D. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *NEJM* **344**, 783–792 (2001).
246. Chapman, P. *et al.* Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *Cancer Res.* 1–10 (2011).
247. Kwak, E. *et al.* Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. *NEJM* **363**, 1693–1703 (2010).
248. Collins, F. S., Morgan, M. & Patrinos, A. The human genome project: lessons from large-scale biology. *Science* (80-.). **300**, 286–290 (2003).

249. NuffieldTrust. Health care spending per person in the UK. (2013). at <<http://www.nuffieldtrust.org.uk/data-and-charts/health-care-spending-person-uk>>
250. Drukker, C. a *et al.* Mammographic screening detects low-risk tumor biology breast cancers. *Breast Cancer Res. Treat.* **144**, 103–11 (2014).
251. Kalager, M., Adami, H., Bretthauer, M. & Tamimi, R. M. Overdiagnosis of Invasive Breast Cancer Due to Mammography Screening: Results From the Norwegian Screening Program. *Ann. Intern. Med.* **156**, 491–99 (2014).
252. Spear, B. B., Heath-chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends Mol. Med.* **7**, 201–204 (2001).
253. Lazarou, J., Pomeranz, B. H. & Corey, P. N. Incidence of Adverse Drug Reactions in Hospitalized Patients A Meta-analysis of Prospective Studies. **279**, (2014).
254. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–54 (2002).
255. Kothari, N. *et al.* Comparison of KRAS mutation analysis of colorectal cancer samples by standard testing and next-generation sequencing. *J Clin Path* (2014).
256. Bettegowda, C., Sausen, M., Leary, R. J. & Kinde, I. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **24**, (2014).
257. Couraud, S. *et al.* Non-invasive diagnosis of actionable mutations by deep sequencing of circulating-free DNA in non-small cell lung cancer: Findings from BioCAST/IFCT-1002. *Clin. Cancer Res.* 1–30 (2014). doi:10.1158/1078-0432.CCR-13-3063
258. Douillard, J. *et al.* Gefitinib Treatment in EGFR Mutated Caucasian NSCLC: Circulating-Free Tumor DNA as a Surrogate for Determination of EGFR Status. *J Thorac Oncol* **9**, 1345–53 (2014).
259. FDA. FDA list of genes with pharmacogenomic indications. (2014). at <<http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm>>
260. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
261. Diegoli, T. M., Farr, M., Cromartie, C., Coble, M. D. & Bille, T. W. An optimized protocol for forensic application of the PreCRTM Repair Mix to multiplex STR amplification of UV-damaged DNA. *Forensic Sci. Int. Genet.* **6**, 498–503 (2012).
262. NEB. PreCR[®] Repair Mix. 9–11 (2014). at <<https://www.neb.com/products/m0309-precr-repair-mix>>
263. Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* **483**, 100–3 (2012).

- 264. McNally, E. Cardiovascular Genetics : Paying Individual Dividends. *Sci. Transl. Med.* **12**, 239–240 (2014).
- 265. Jamal-Hanjani, M. *et al.* Tracking Genomic Cancer Evolution for Precision Medicine: The Lung TRACERx Study. *PLoS Biol.* **12**, e1001906 (2014).
- 266. Schrappe, M. Minimal residual disease : optimal methods , timing , and clinical relevance for an individual patient. 21–24 (2012).
- 267. Hourigan, C. S. & Karp, J. E. Minimal residual disease in acute myeloid leukaemia. *Nat. Rev. Clin. Oncol.* **10**, 460–71 (2013).

Appendix 1: Letters of support

Letters of support from co-authors of the manuccripts submitted as part of this PhD by Publication are reproduced here. Copies of all published work are self-contained in the Appendix 2 submitted separately to this thesis.

UEA, Norwich
NR4 7TJ England

Telephone 01603 456161
Direct dial 01603 591167
Email b.jennings@uea.ac.uk



Dear Colleagues,

Re. James Hadfield's contribution to the study

Jennings, B. A., Hadfield, J., Worsley, S. D., Girling, A. & Willis, G. A differential PCR assay for the detection of c-erbB 2 amplification used in a prospective study of breast cancer. Mol. Pathol. 50, 254–256 (1997).

The study described in this publication was initiated in 1994 as part of a wider programme of work within the pathology department of the Norfolk and Norwich Hospital (now NNUH). The aims were to establish molecular techniques that could be more sensitive and specific than traditional histological and cytogenetic techniques for cancer diagnostics. We explored many techniques for analysing oncogenes, their transcripts, and their protein products. This work was conducted before automated systems existed for Real Time PCR and before NICE guidelines were available to recommend particular companion diagnostics, both of which are now incorporated into molecular pathology services at NNUH.

The focus of the study that James contributed to was to test the analytic validity and clinical validity of the assay that we designed. James collected and analysed the vast majority of the genotype data and the data needed to appraise the analytic validity of the test; with careful and systematic use of controls and dilution series of DNA from cell lines with known copy numbers of the *ERBB2* locus. James also contributed to the first and last drafts of the manuscript including the literature reviews. The sample collection preceded James' tenure in the lab and some of the phenotype data was collected in the year after he left, so overall I would estimate that James' contribution comprised about 25 % of the published study. However, James contributed 75 % of the data analysis and interpretation of analytic validity. Given that this was James' first post-graduate post and that the grant-funded appointment was only for 10 months, he made a remarkable and invaluable contribution to this study. Furthermore, he has continued to build on and pioneer techniques designed for efficient analysis of human nucleic acids; to appraise and develop genome study design; and to generate research data and contribute to their translation into genomic tests with potential clinical utility.

Yours sincerely,

A handwritten signature in purple ink, which appears to read 'Barbara Jennings', is positioned above the typed name.

Barbara Jennings PhD
Senior Lecturer in Molecular Medicine
Norwich Medical School
Faculty of Medicine and Health Sciences.

London, 28th April 2014

To whom it may concern,

this letter is to confirm James Hadfield's contribution to the following publication:

- Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., & Odom, D. T. (2009). ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. *Methods*, 48(3), 240-248

I was the lead author on the above mentioned paper, which has become one of the most cited methods in this field, being referenced over 130 times to date. James was invited to submit a manuscript to the journal *Methods* and we worked together on the final paper. At the time of publication there was no detailed method for Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) available. ChIP-seq has become a routine research tool and James was involved at all stages of the manuscript preparation. His technical experience in next-generation sequencing and intellectual contribution was invaluable to making this methods paper focus on all aspects of the experiment not just the sequencing library preparation. He personally developed some of the quality control tools that are now routine use in several institutes. The protocol and methods published in this paper have been used in tens of published papers and hundreds of experiments at the Cambridge Institute. James continues to support researchers in experimental design for ChIP-seq and other experiments, which has been acknowledged in almost 100 papers, some of which in *Nature*, *Science*, *Cell* and other high-impact journals.

This paper has made an important contribution to the field and it was very enjoyable working with James on its preparation and publication. I am fully supportive of his Ph.D. application and am happy to provide any further feedback on request.

Yours faithfully,



Dominic Schmidt, Ph.D.
Partner

Supported by
wellcometrust

Syncona Partners LLP, 215 Euston Road, London NW1 2BE, UK
T +44 (0)20 7611 2031 F +44 (0)20 7611 2032 E contact@synconapartners.com synconapartners.com

Syncona Partners LLP is a limited liability partnership registered in England and Wales under registration number OC377978. Registered office: 215 Euston Road, London NW1 2BE, UK.

Michael D. Wilson, PhD
Canada Research Chair in Comparative Genomics
Scientist (SickKids), Assistant Professor (U of T)

Department of Molecular Genetics
University of Toronto

Program in Genetics and Genome Biology
SickKids Research Institute
Peter Gilgan Centre for Research and Learning
686 Bay Street | Room 14.9713
Toronto M5G 0A4
Canada
Tel: (416) 813-7654 ext 328699
michael.wilson@sickkids.ca

28-March-2014

To whom it may concern,

This letter is written to describe James Hadfield's contribution the following scientific publication:

Schmidt D, **Wilson MD**, Spyrou C, Brown GD, Hadfield J, Odom DT. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*. 2009 Jul;48(3):240-8.

This comprehensive methods has become a go-to resource for ChIP-seq methods that has been cited 130 times since its publication. The existence of this publication can be directly attributed to James Hadfield who suggested we work together to create a comprehensive resource for ChIP-seq at a time when no such resources existed. At the time of writing no such detailed description of the method from tissue preparation through to next generation sequencing existed. James played a leadership role on this paper -- in addition to its conception, he was involved with all stages of the manuscript preparation. Furthermore, James's direct experience with users of his genomics core gave us a unique perspective regarding the details to include in the protocol and the manner how we described them. I believe that this perspective helped make the manuscript more accessible and that it can explain the high number of citations the article has obtained since its publication. Five years later I still give this manuscript to my students or colleagues who want to learn ChIP-seq. In addition to writing, James played an instrumental role in sequencing and performing quality control hundreds of ChIP-seq libraries that were generated using this method (these have been acknowledged in several publications as well, including Schmidt, Wilson, Ballester et al. *Science* 2010). If you require any further information regarding James's involvement in this publication please do not hesitate to contact me.

Sincerely,



Michael D. Wilson

2014 April 2

Dr. Gordon Brown
Senior Bioinformatician
CRUK Cambridge Institute
Robinson Way
Cambridge CB2 0RE

To whom it may concern:

This letter describes James Hadfield's contributions to the *Methods* article:

Dominic Schmidt, Michael D. Wilson, Christiana Spyrou, Gordon D. Brown, James Hadfield, Duncan T. Odom. "ChIP-seq: Using high-throughput sequencing to discover protein-DNA Interactions", *Methods* 48:240—248 (2009).

James made several significant contributions to the article. Perhaps his most significant is that he was the individual originally invited by the editors of *Methods* to contribute the manuscript to the journal, due to his prominence in the high-throughput sequencing world. He played a substantial role in the preparation of the manuscript, overseeing the efforts of the other authors, in addition to carrying out the sequencing and much of the library quality control. He was the primary author of the sequencing section.

The article itself remains the most complete, detailed description of the ChIP-seq protocol available. It has been cited more than 130 times, in fields as diverse as cancer research, agriculture, and evolution. As ChIP-seq has become overwhelmingly the dominant technique for studying transcription factor binding and DNA/histone modifications genome-wide, the article's value to the field is very high.

Please feel free to contact me if you require any further information.

Regards,


Gordon Brown

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE
Tel: +44 (0)1223 769 500

www.cruk.cam.ac.uk

26 March 2014

Dr Duncan T Odom

Group Leader
Regulatory Systems Biology Laboratory
Li Ka Shing Centre
Robinson Way
Cambridge CB2 0RE

To whom it may concern:

I am writing today as the corresponding author on the following publication:

D. Schmidt, M. Wilson, C. Spyrou, G. Brown, **J. Hadfield**, D. T. Odom. "ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions," *Methods* **48** (2009) 240-248.

Citations: 130 total, as of 25 March 2014.

At the time of this manuscript's publication, I was a tenure-track junior group leader at the University of Cambridge / Cancer Research UK – Cambridge Institute, where James Hadfield has served as the genomic core director for the last seven and a half years.

James Hadfield provided both substantial intellectual and technical input into this manuscript. He assisted with the initial experimental design, confirmed the quality of the sequencing libraries and ChIP experiments, and helped write the paper. His contribution substantially strengthened this paper, to the point where we discussed the possibility of his being a co-corresponding author. The final manuscript has become an exceptionally well cited paper, particularly for a methodology report.

I consider James to be a strongest possible candidate for a PhD, and support his application enthusiastically. I am happy to provide any further information, as requested.

Warmest regards,



Re: Curtis et al, 2009

Dear Sir/Madam,

I am writing with reference to the publication: The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*. 2009 Dec 8;10:588.

As co-author of the above paper, and as Team leader for the sample and data collection for the Metabric project, I can confirm that Mr. James Hadfield , Core Facility Manager for Genomics at the CRUK Cambridge Institute had played an important role in the project. James, not only participated in the experimental design, the different platform tested but the actual data acquisition as well.

The work done in this project was important in determining the platform that was chosen for the analysis of 2000 primary breast tumours which resulted in our seminal paper published in Nature in 2012.

Yours sincerely

Suet-Feung Chin PhD (Cantab)
Associate Scientist
Functional Breast Cancer Genomics Group
CRUK Cambridge Research Institute
Li Ka Shing Centre
Robinson Way
Cambridge
CB2 0RE
UK

Dr Anna Git

Cancer Research UK Cambridge Institute &
University of Cambridge, Department of Oncology
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE, UK
Tel: +44 (1223) 769728
Email: Anna.Git@cruk.cam.ac.uk

To: all whom it may concern

Dear Sir or Madam,

In support of James Hadfield's Ph.D. by Publication I am writing regarding the following study, in which I acted as lead investigator, and the resulting paper of which I am the first and corresponding author:

"Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression". Git A., Dvinge H., Salmon-Divon M., Osborne M., Kutter C., Hadfield L., Bertone P. and Caldas C. *RNA* 16(5):991-1006.

At the end of 2008, as part of the METABRIC project, our laboratory was planning to embark on a microRNA profiling study of over 1,000 primary breast tumours collected by an international consortium. While the importance of microRNAs in normal and tumour biology was becoming evident, the relevant genome-wide technologies and their analyses were in their infancy. We therefore teamed up with the James Hadfield and initiated a thorough comparison between several existing methodologies. James' knowledge of current technological developments and his experience with microarray platforms were instrumental in the design of the study. He then proceeded to lead the logistical aspects of the study, liaising with manufacturers and dedicating facility resources to the execution of the study.

The resulting paper remained the most highly accessed *RNA* manuscript for a year and has been cited almost 200 times (Google Scholar) to date. It has been widely acknowledged as the best of its sort (informal communication) and has set our team as international experts on microRNA analysis. Furthermore, the platform chosen as a result of our preliminary study has been later used to examine the microRNA profiles of 1,300 breast tumours, and their analysis was published in *Nature* (Dvinge *et al.* 2013). I have no doubt that the meticulous planning of our preliminary study was critical in obtaining reliable data from precious clinical samples.

It is difficult to enumerate the exact contribution of individuals to what was very much a team effort. In terms of time and effort, I would estimate that James contributed 5%, but his advice and experience steered us toward a robust experimental layout and away from common pitfalls and were thus crucial to the publication.

Please do not hesitate to contact me for any further information or clarification.

Yours faithfully,

Anna



UNIVERSITY OF
CAMBRIDGE

25th of March 2014

University of East Anglia
Norwich Research Park
NR4 7TJ Norwich, Norfolk
United Kingdom

Claudia Kutter
University of Cambridge
Cancer Research UK Cambridge Institute
Li Ka Shing Centre
Robinson Way
Cambridge CB2 0RE
United Kingdom

Tel: +44 (0) 1223 769623
Fax: +44 (0) 1223 769881
Skype: claudia.kutter
Email: claudia.kutter@cruk.cam.ac.uk

Re: Git *et al.*, 2010

Dear Sir and Madam,

James Hadfield from the Cancer Research UK Cambridge Institute asked me to provide a letter of support for his PhD by publication. It is with great pleasure to act James' request to commend his contribution to the following paper:

Anna Git, Heidi Dvinge, Mali Salmon-Divon, Michelle Osborne, Claudia Kutter, **James Hadfield**, Paul Bertone, and Carlos Caldas: "*Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression*"; RNA, Vol. 16, No. 5, 2010

The focus of this study is centered on a specific class of short regulatory RNAs, called microRNAs (miRNAs), which regulate protein-coding gene expression in a sophisticated manner and thus, control crucial biological processes such as carcinogenesis. These miRNAs hold great potential as diagnostic marker and therapeutic agent. Hence, correct measurement of the cell-type specific expression of miRNAs is essential for any future applications. Multiple miRNA profiling platforms had been described when this project started. Each of them harbored potential biases, making it difficult to distinguish signal from noise. This study addressed this challenge by systematical profiling and comparison of miRNA expression of three well-defined samples across six commercially available miRNA microarray platforms, Illumina next-generation sequencing (NGS) and standard quantitative PCR.

To date, this paper has been highly accessed and cited (185 times), which is on a scale of top-tier journals (such as Cell, Science and Nature) and represents one of the landmark papers in this field.

As a co-author on the above-mentioned paper, I interacted strongly with James from the start until the completion of this project. James contributed tremendously in the experimental design, planning and execution of this research project. He helped in the analysis of the data, interpretation of the results and preparation of this work for publication. From the beginning, James stressed the importance of ensuing data reproducibility, robustness in the data collection and downstream data mining. James identified commercially available platforms that work in different ways to increase technical variability, which helped enormously in the data analysis to distinguish real signal from noise. Including the profiling of miRNAs by using NGS was James' suggestion right from the start. There was some reluctance and skepticism by other project members because this sequencing technology was new in the building at this point and the methods (for both sample preparation and sequencing) needed to be established. James personally ensured that these efforts were pushed ahead. It showed his forward thinking in which direction the field will develop and what questions one needs to anticipate. Retrospectively, adding these NGS data was essential if not vital to our study. Other publications refer to this particular result, demonstrating that these data contribute to the remaining scientific interest in this study, which is surprising in this research area because of rapid changes in technology and fast turnover of sequencing data.

As I recall, this work attracted a lot of attention within the scientific "-omics" community and James presented parts of this work at many specialized meetings, workshops and conferences.

Yours sincerely,

Claudia Kutter



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

30th April 2014

Dr Andy G Lynch
Statistics and Computational Biology Group
Cancer Research UK Cambridge Institute
University of Cambridge

Dear Sir/Madam

I am writing regarding the publication:

Lynch AG, Hadfield J, Dunning MJ, Osborne M, Thorne NP and Tavaré S (2010) "The cost of reducing starting RNA quantity for Illumina BeadArrays: A bead-level dilution experiment", BMC Genomics 11:540

This work rose out of a research need to deal with important clinical samples that fail to yield the quantities of nucleic acids required by standard protocols, and moreover to do this in a manner that allows them to be directly compared to other samples for which the standard protocols were appropriate. This provided the ideal basis for a collaboration between James Hadfield who has an interest in pushing the limits of technologies, and myself and colleagues with our interests in maximizing the amount of information that can be leveraged from apparently 'sub-optimal' experiments.

The design process for the experiment was complex as the field was changing rapidly, and our ambitions became greater over time. James made great contributions to the experimental design discussions over many months, and his knowledge both of the technology, and of other initiatives to assess the technology, was invaluable.

James also contributed to the write-up and presentation of the results, ensuring that we retained relevance to the broadest possible community, and he also presented this work at the 2010 Illumina User Group Meeting.

Yours sincerely,

Dr Andy G. Lynch

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE
Tel: +44 (0)1223 769 837

www.cruk.cam.ac.uk



UNIVERSITY OF
CAMBRIDGE

26th of March 2014

University of East Anglia
Norwich Research Park
NR4 7TJ Norwich, Norfolk
United Kingdom

Sarah Aldridge
University of Cambridge
Cancer Research UK Cambridge Institute
Li Ka Shing Centre
Robinson Way
Cambridge CB2 0RE
United Kingdom

Email: sarah.aldridge@cruk.cam.ac.uk

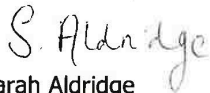
Dear Sir/Madam,

I am writing with reference to James Hadfield's PhD by publication and in particular to the following invited book chapter:

Sarah Aldridge and James Hadfield, Introduction to miRNA profiling technologies and cross-platform comparison, Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols, Methods in Molecular Biology, Vol 822, 2012.

James is (and was at the time we co-authored this book chapter) a leading expert in the use of genomic technologies and was invited to write the book chapter because of this. As I recall, at the time of writing, the use of next generation sequencing for analysis of microRNAs was a new and novel technique that James devoted a significant amount of time and energy into researching, optimising and communicating to the Genomics community. This was of significant benefit to the CRUK Cambridge Institute and the Genomics community as a whole. James took the lead on writing this book chapter and contributed at least 50% (if not more) to the process from start to finish.

Yours sincerely,


Sarah Aldridge

February 4th, 2015

To Whom it May Concern,

I am writing in regards to James Hadfield's Ph.D. by publication and his contributions to the following publications:

Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin SF, Brenton J, Tavaré S, & Caldas C. The pitfalls of platform comparison: A comparison of DNA copy-number platforms. *BMC Genomics*, 10:588, 2009.

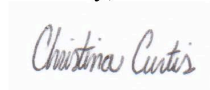
Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, METABRIC Group, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Borrensens-Dale AL, Brenton JD, Tavaré S, Caldas C & Aparicio S. The integrative genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346-52, 2012.

In the 2009 Curtis *et al.* paper we performed a careful assessment of the inference of copy number alterations in primary breast tumors and corresponding matched normals, established cancer cell lines and HapMap individuals. In particular, we profiled these samples on four leading microarray platforms, namely the Affymetrix Genome-wide SNP 5.0 array, Agilent High-Density CGH Human 244A array, Illumina HumanCNV370-Duo DNA Analysis BeadChip, and the Nimblegen 385 K oligonucleotide array and performed a theoretical assessment of the reproducibility, noise, and sensitivity. James was involved in various aspects of this study ranging from platform selection to sample hybridization. Not only was this study important to characterizing platform performance, but it also contributed to our choice of copy number profiling platform for the METABRIC study.

In the Curtis *et al.* 2012 Nature paper, we exploited integrative statistical approaches to mine multiple genomic data types, revealing novel subtypes of breast cancer with distinct clinical outcomes and subtype-specific driver genes. This landmark study has become a resource to the community and redefined the molecular map of breast cancer. James played an instrumental role in the Genome and Transcriptome Characterization Centre and oversaw the transcriptional profiling of all tumors and matched normals. He also contributed to the study and implementation, and hence contributed both intellectually and technically.

Feel free to contact me should you have any questions.

Sincerely,



Christina Curtis, Ph.D., MSc.
Assistant Professor of Medicine & Genetics
Co-Director, Molecular Tumor Board
Stanford Cancer Institute
Stanford University School of Medicine

Lorry Lokey Building, Suite G2120C
265 Campus Drive | Stanford, CA 94305-5456

Re: Curtis et al, 2012

Dear Sir/Madam,

I am writing with reference to the publication: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups published in *Nature*. 2012 Apr 18;486(7403):346-52.

As joint first author of the above paper, and as Team leader for the sample and data collection, I can confirm that Mr. James Hadfield , Core Facility Manager for Genomics at the CRUK Cambridge Institute had played an important role in the project. In particular, the RNA expression data was generated in the facility and James, not only participated in the experimental design but the actual data acquisition as well.

This publication is a seminal paper in further our understanding of breast cancers.

Yours sincerely

Suet-Feung Chin PhD (Cantab)
Associate Scientist
Functional Breast Cancer Genomics Group
CRUK Cambridge Research Institute
Li Ka Shing Centre
Robinson Way
Cambridge
CB2 0RE
UK

7th January 2015

To whom it may concern:

James Hadfield contributed to the following published work involving my group:

Curtis et al 2009: The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC Genomics 2009; 10:588-611

Git et al: Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA 2010; 16:991-1006

Curtis et al 2012: The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. Nature 2012; 486:346-52

Forsheew et al: Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci. Transl. Med. 2012; 4(136): 136ra68

Murtaza et al: Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 2013; 497:108-12

James as part of his role of Genomics Facility manager ensured that we were informed about the latest technologies available that fitted our research. For instance, he was involved in the discussions about what microarray platforms to test for two microarray comparison projects resulting in two publications ie Curtis et al 2009 and Git et al 2010. He was pivotal in working with Illumina on the design of the HT12 gene expression array which reduced batch effects significantly and also lowered costs two fold. Both of the above mentioned projects were performed as pilot experiments to choose the most suitable platforms to use for the METABRIC study. This resulted in Curtis et al 2012, which is a seminal paper that described novel molecular sub-groups of breast cancer. Over 2000 breast cancer samples were used, and half of these had both DNA and RNA extracted, quantified and normalised by staff in the genomics Core Facility. James coordinated the processing of almost 3000 HT12 gene expression arrays in the Genomics core, and also worked with service providers to arrange the processing of over 2500 Affymetrix SNP6.0 genotyping arrays for copy number and LOH analysis.

James contributed to the early development of amplicon sequencing using the Fluidigm Access Array. He was involved in the application of this to circulating tumour DNA and his experience of next-generation sequencing technologies influenced the early discussions about different sequencing methods and their potential strengths and weaknesses. He helped in the optimisation and implementation of the TAM-seq method. James also worked on the evaluation of the ThruPLEX technology used to prepare ctDNA libraries for exome sequencing of ctDNA published in Murtaza et al 2013.



Professor Carlos Caldas
Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE
Tel: +44 (0)1223 769650

www.cruk.cam.ac.uk



Dr Tim Forshaw
UCL Cancer Institute
UCL
72 Huntley St
London
WC1E 6DD
t.forshaw@ucl.ac.uk

To whom it may concern

Dear Sir/Madam,

I write with regards to the following publication and James Hadfield's contribution to this work:

Forshaw T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D, **Hadfield J**, May AP, Caldas C, Brenton JD, Rosenfeld N. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. **Sci Transl Med**. 2012 May 30;4(136):136ra68.

This study was designed, completed and published while James Hadfield and I worked together at the Cancer Research UK Cambridge Institute (CRUKCI). In this study we describe a novel approach to screen for mutant DNA, released by solid tumours into circulation. To our knowledge this is the first next generation sequencing study to demonstrate detecting de novo solid tumour somatic mutations through unbiased plasma sequencing. This and subsequent manuscripts (on which he also contributed) have made significant impact on this field. I am personally aware of a number of groups throughout the world testing these approaches on different types of cancer.

James initiated the amplicon based sequencing at CRUKCI on which this was based. He was heavily involved with discussions as to how different sequencing methods work and the potential strengths and weaknesses of each. He oversaw the sequencing of the libraries used in this paper. Finally James was involved with writing this manuscript.

Please feel free to contact me for any more information on James's contribution.

Yours faithfully,

Dr Tim Forshaw
Group leader
UCL Cancer Institute

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre
Robinson Way
Cambridge
CB2 0RE

28 March 2014

Dear Sir/Madam,

I am writing with reference to the following publication:

T. Forsheew, M. Murtaza, C. Parkinson, D. Gale, D. W. Y. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld, Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* 4, 136ra68 (2012).

At the time this paper was written, James Hadfield and I were working at the Cancer Research Cambridge Institute (Formerly known as Cancer Research UK Cambridge Research Institute). At that time, James were the Head of the Genomic Core facility of the institute.

The research describes a novel method, termed Tagged-Amplicon deep Sequencing (TAM-Seq), for detection and quantification of tumour-specific somatic point mutations in plasma of cancer patients. The main advantage of TAM-Seq over technologies available at the time of the publication is that it allows robust and accurate measurement of tumour-specific DNA across sizable genomic regions in blood plasma in a high-throughput and cost-effective manner. It opened up possibilities for large-scale validation study to investigate the clinical utility of circulating tumour DNA as a non-invasive monitoring tool for cancer management.

James' major contributions to the research were in various aspects of the method development, in particular negotiating with the company to get access to the system for optimal PCR-based techniques to amplify rare DNA molecules, optimising the implementation of sample-specific identifiers to improve throughput, and improving the experimental procedures to optimise turn-over time. James and his team performed the sequencing analysis. James assisted in writing the manuscript and approved the final version.

Yours

Dana WY Tsui

Postdoctoral Research Fellow
Cancer Research UK Cambridge Institute

Dana.tsui@cruk.cam.ac.uk
+44(0)1223769736



March 25, 2014

To Whom It May Concern:

**Re: Contributions by James Hadfield as co-author of Forshew et al. Science
Translational Medicine 2012.**

Forshew et al. described an approach for massively parallel sequencing of fragmented low-molecular weight circulating tumor-specific DNA in plasma to identify low-abundance somatic mutations in blood of cancer patients. I was co-first author on this manuscript and worked with James Hadfield at the Cancer Research UK Cambridge Institute in my capacity as a doctoral candidate.

James Hadfield contributed to our recognition of amplicon-based sequencing as a viable strategy for sequencing gene-sized regions from degraded DNA fragments. He further facilitated the identification and procurement of a microfluidic platform (Fluidigm's Access Array) that could enable the sequencing approach. His knowledge and insight into sequencing chemistries and methods was useful in designing and planning experiments using chimeric primers for highly multiplexed sequencing. He directly facilitated the implementation of these experiments and generation of sequencing data. Finally, he was particularly useful during preparation of the manuscript in comparing and distinguishing our approach from competing technologies and helped emphasize key improvements over previously published methods.

This manuscript made novel technological improvements in ctDNA sequencing, enabling downstream applications in clinical research studies. Dawson et al. used our approach and demonstrated ctDNA was a superior biomarker for monitoring tumor burden in metastatic breast cancer patients than circulating tumor cells and CA15-3 (Dawson et al. NEJM 2013). Deep sequencing of ctDNA achieving thousands-fold coverage, as described in our paper, has been proposed as a tool for evaluating therapeutic selection and clonal evolution (Aparicio et al. NEJM 2013). In less than 2 years, our paper has been cited 84 times (Google Scholar) as a useful strategy for non-invasive molecular solid cancer diagnostics and as a

technical advance in amplicon-based targeted sequencing and low abundance mutation detection.

In summary, Forshew et al. adapted cutting edge technologies in genomics to enable analysis of ctDNA that was previously difficult at this scale. James made key intellectual and practical contributions to this work.

Please feel free to contact me if there is any further information I may provide.

Yours sincerely,



March 25, 2014

Muhammed Murtaza, MBBS

Research Assistant Professor

Co-Leader, Program in Circulating Nucleic Acids,

Translational Genomics Research Institute (TGen),

Phoenix, AZ, USA.

email: mmurtaza@tgen.org

phone: +1 602 343 8497

References:

Forshew et al. **Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA.** *Sci. Transl. Med.* 4, 136ra68 (2012).

Dawson et al. **Analysis of circulating tumor DNA to monitor metastatic breast cancer.** *N Engl J Med.* 2013 Mar 28;368(13):1199-209.

Aparicio S. and Caldas C. **The implications of clonal genome evolution for cancer medicine.** *N Engl J Med.* 2013 Feb 28; 368(9):842-851

Peter MacCallum Cancer Centre
St Andrews Place
East Melbourne Victoria
Postal Address
Locked Bag 1 A'Beckett Street
Victoria 8006 Australia
Phone +61 3 9656 1111
Fax +61 3 9656 1400
ABN 42 100 504 883
www.petermac.org

Locations
East Melbourne
Bendigo
Box Hill
Moorabbin
Sunshine



30th March 2014

To whom it may concern,

Re: James Hadfield
Letter of Support
Contribution as co-author of Forshew et. al. Science Translational Medicine 2012

Forshew et. al. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA" was published in Science Translational Medicine in 2012. The manuscript, for which I was a co-author, detailed the development of a targeted sequencing approach (Tagged-Amplicon Sequencing (Tam-Seq)) for the detection of tumour mutations in plasma DNA.

The benefit of measuring circulating nucleic acids for biomarker applications in cancer has only begun to be explored in the last decade. Higher levels of circulating nucleic acids are identified in cancer patients compared to healthy controls due to the presence of circulating DNA containing tumour-specific sequences (ctDNA) that harbor the somatic mutations found in a patient's tumour. The analysis of ctDNA is challenging and requires highly sensitive techniques due to the small fraction of tumour specific DNA present within background levels of total genomic DNA. Tagged-Amplicon Sequencing (TAM-Seq) is a method which uses a microfluidic device (FluidigmTM) to generate multiplexed libraries of barcoded short amplicons allowing targeted sequencing of plasma. TAM-Seq enables sequencing of ~10Kb of the genome from a few copies of fragmented DNA at 10,000 fold coverage, using amplicon panels that are flexible and easy to expand. Prior to the development of the Tam-Seq methodology, various methods had been optimized to detect rare mutations at individual loci, but techniques to identify mutations at low allele fractions across sizeable genomic regions had been more challenging.

James was closely involved in the development of the TAM-Seq methodology. When faced with the challenges of sequencing circulating DNA which is highly fragmented, James facilitated the procurement of the FluidigmTM microfluidic platform and the application of an amplicon-based sequencing approach. His extensive experience in sequencing technologies provided valuable assistance in the optimization and implementation of the TAM-Seq methodology, and he provided critical input during the preparation of the manuscript.

This study has opened up new opportunities to develop ctDNA as a personalised biomarker in various solid malignancies. In particular, the TAM-Seq methodology was applied in Dawson et. al. "Analysis of circulating DNA to monitor metastatic breast cancer" NEJM 2013, which demonstrated the superiority of ctDNA as a biomarker for disease monitoring in breast cancer.

Peter MacCallum Cancer Centre

St Andrews Place
East Melbourne Victoria
Postal Address
Locked Bag 1 A'Beckett Street
Victoria 8006 Australia
Phone +61 3 9656 1111
Fax +61 3 9656 1400
ABN 42 100 504 883

www.petermac.org

Locations

East Melbourne
Bendigo
Box Hill
Moorabbin
Sunshine



These studies have provided a paradigm for the use of ctDNA for molecular disease monitoring in solid malignancies, a strategy that has not previously been possible. Through his involvement in the development of the TAM-Seq methodology, James has made a significant contribution to this work.

Yours sincerely,

Dr Sarah-Jane Dawson (MBBS, FRACP, PhD)
Consultant Oncologist &
Group Leader, Molecular Biomarkers and Translational Genomics Laboratory
Peter MacCallum Cancer Centre
St Andrew's Place, East Melbourne
Victoria 3002, AUSTRALIA
Phone +61 3 9656 3728
E-mail Sarah-Jane.Dawson@petermac.org

26th March 2014

Dear Sir/Madam,

I am pleased to acknowledge James Hadfield's contribution to the work described in the manuscript below, on which I am a co-author.

Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA
Forshew *et al. Sci Transl Med* 4, 136ra68 (2012)

James and I worked together from 2009 to 2012 while I was leading R&D efforts at Fluidigm to develop products for high-throughput amplicon sequencing.

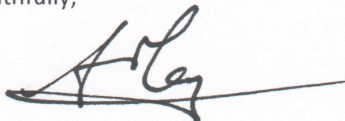
Early in 2009, the majority of the sequencing community was focused on exome sequencing and array capture as a method for enriching selected regions from genomic samples. At Fluidigm, we had developed a system (the Access Array system) and protocol for amplifying multiple PCR products from many samples in parallel and attaching sequencing tags and barcodes for use with the 454 Sequencing system. James was one of the few people in the field to recognize that, if applied to the much higher throughput Illumina system, our approach would enable the sequencing of hundreds or thousands of samples in parallel. James established a collaboration between my group and his lab at Cancer Research UK to demonstrate the feasibility of high-throughput amplicon sequencing on the Illumina system, and we quickly established that we were able to detect mutations in multiple loci from multiple samples in parallel.

James was quick to recognize that this could be applied in many settings within cancer genomics, and introduced me to Nitzan Rosenfeld and James Brenton (communicating authors on the manuscript) to explore the possibility of using the system for amplification from FFPE samples and circulating DNA from the plasma of ovarian cancer patients. James was involved in the initial study design, and worked with Tim Forshew (the first author on the paper) to demonstrate feasibility of the study. His lab was then further involved in running samples to complete the study. James contributed to writing the manuscript and collected additional data during rounds of submission, review and resubmission until the study was finally accepted for publication.

The manuscript itself was one of the first examples to demonstrate the detection of mutations in multiple regions from circulating plasma DNA, was the first manuscript to demonstrate the ability to follow genome sequencing with rapid amplicon sequencing as a method of monitoring patient progress during treatment, and laid the foundations for multiple subsequent manuscripts demonstrating the ability to sequence circulating cell-free DNA as a means to monitor tumours before, during and after treatment.

James' early recognition of the need to develop methods for high sample-throughput, as well as high sample content, was a key driver in initiating the collaboration that ultimately led to this publication, which has proved foundational in the use of circulating tumour DNA for patient diagnosis and monitoring.

Yours faithfully,



Andrew May, D. Phil.
Chief Scientific Officer,
Caribou Biosciences, Inc.



Illumina, Inc.
5200 Illumina Way
San Diego, CA 92122
tel 858.202.4500
fax 858.202.4545
www.illumina.com

March 26th, 2014

To whom it may concern,

As one of the co-authors of the following paper, I can confirm that James Hadfield has played a critical role in bringing this paper to fruition:

T. Forshaw, M. Murtaza, C. Parkinson, D. Gale, D. W. Y. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012)

James was instrumental in the development of the method described in the paper (TAm-Seq) which demonstrated the feasibility of simultaneously tracking mutations in multiple genes in circulating tumor DNA. Prior to publication of this method, researchers commonly applied PCR-based methods to detect predefined mutations in these liquid biopsies. The major advantages of TAm-Seq are the detection of de novo mutations in genes that lack hotspot mutations and the high levels of multiplexing. The paper was picked up by several news agencies including The Telegraph, Daily Mail, and LA Times emphasizing the importance of its findings.

I collaborated with James while on the RnD team at Fluidigm where I worked on the development of the Access Array System. It was due to James' insight and guidance that we realized the potential application of this system in high throughput screening of cancer samples. He subsequently initiated the multi group collaboration that led to the published research study. He was involved in designing the feasibility experiments, in sequencing the samples, and writing of the manuscript. Additionally, James brought a wealth of experience in Next Generation Sequencing (NGS) to the table, without which this paper would not have achieved the high levels of data quality.

Yours Sincerely,

A handwritten signature in blue ink, appearing to read "Fiona Kaper", with a stylized flourish at the end.

Fiona Kaper, PhD
Staff Scientist
Illumina, Inc.



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

8 January 2015

To whom it may concern,

Re: Contribution of James Hadfield to publication by
Forshaw et al., Science Translational Medicine 2012

Dear Sir or Madam:

In reference to the following publication, of which I am a corresponding and senior author:

T. Forshaw, M. Murtaza, C. Parkinson, D. Gale, D. W. Y. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld, Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* 4, 136ra68 (2012)

I would like to write to highlight the contributions of James Hadfield.

At the time this work was developed and performed and the paper was written, I was a Junior Group Leader at the Cambridge Research Institute (now the Cancer Research UK Cambridge Institute), where James Hadfield was head of the Genomics core facility.

The work described in this paper was developed, to a significant part, thanks to contributions and input by James, through his capacity as head of the Genomics core facility. James was involved in suggesting and discussing approaches based on amplicon sequencing, identifying and securing access to technologies used, and overseeing sequencing work carried out in the core facility he managed. James was involved in discussing and interpreting data, results and technical issues, and aided in writing, reviewing and approving the manuscript.

This paper, published in 2012, is highly influential in the development of the circulating tumour DNA field. As evidence of its impact, it was reviewed and covered by news items in high-circulating newspapers such as the Telegraph and Daily Mail, and in the short time that has elapsed has already been cited more than 150 times in the professional literature. The methods described in the paper have been the basis of significant further work in the field, including work leading to highly influential publications such as a paper by Dawson, Tsui et al. published in the New England Journal of Medicine in 2013 (DOI: 10.1056/NEJMoa1213261).

Sincerely,

Nitzan Rosenfeld

Senior Group Leader

Cancer Research UK Cambridge Institute, University of Cambridge

Tel: 01223-769769, Fax: 01223-769510, e-mail: nitzan.rosenfeld@cruk.cam.ac.uk

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE

Tel: +44 (0)1223 769 500

www.cruk.cam.ac.uk



March 25, 2014

To Whom It May Concern:

Re: Contributions by James Hadfield as co-author of Murtaza et al. Nature 2013.

Murtaza et al. described the use of low-input whole-exome sequencing of circulating tumor DNA (ctDNA) for non-invasive monitoring of clonal evolution in solid cancers and for identification of novel drivers of acquired therapeutic resistance. I led the study as co-first author of the manuscript and worked with James Hadfield at the Cancer Research UK Cambridge Institute in my capacity as a doctoral candidate.

When we conceived the use of exome sequencing of ctDNA to assess clonal evolution, one of the key challenges was low amount of input DNA available in a plasma sample. With his experience and insight in massively parallel sequencing, James was instrumental in evaluating available and upcoming solutions to make exome sequencing libraries from a few nanograms of degraded plasma DNA. After failures with early experiments, James facilitated evaluation and procurement of a commercially available solution from Rubicon Genomics (ThruPLEX) that enables preparation of high diversity sequencing libraries from as little as 2ng DNA in our experience. He further provided useful input to the manuscript when describing the potential impact and limitations of our approach.

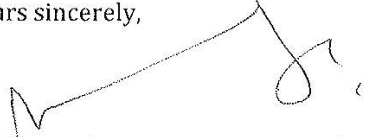
This was a groundbreaking study that demonstrated non-invasive, in vivo in human assessment of clonal evolution in solid cancers using ctDNA sequencing for the first time. Our results triggered a specially invited comment from opinion leaders in clonal evolution and ctDNA analysis (Pantel, Diaz and Polyak. Nature Medicine 2013). Follow up studies using ctDNA exome sequencing aimed at discovery of novel drivers of acquired resistance within clinical trials are currently being pursued by a number of groups. In less than a year, our paper has been cited 73 times (Google Scholar) as a novel approach, potentially revolutionizing assessment of acquired resistance in non-hematological metastatic cancers.

In summary, Murtaza et al. relied on advances in genomic sequencing protocols that could work with low amounts of starting material. We applied these methods to work with ctDNA

to address clonal evolution and acquired resistance. Our collaboration with James was instrumental in pursuing and in implementing the right solutions required to perform this work.

Please feel free to contact me if there is any further information I may provide.

Yours sincerely,



March 25, 2014

Muhammed Murtaza, MBBS

Research Assistant Professor

Co-Leader, Program in Circulating Nucleic Acids,

Translational Genomics Research Institute (TGen),

Phoenix, AZ, USA.

email: mmurtaza@tgen.org

phone: +1 602 343 8497

References:

Pantel K, Diaz LA and Polyak K. **Tracking tumor resistance using 'liquid biopsies'**. *Nat Med*. 2013 Jun;19(6):676-7.

Murtaza et al. **Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA**. *Nature*. 2013 May 2;497(7447):108-12.

Peter MacCallum Cancer Centre
St Andrews Place
East Melbourne Victoria
Postal Address
Locked Bag 1 A'Beckett Street
Victoria 8006 Australia
Phone +61 3 9656 1111
Fax +61 3 9656 1400
ABN 42 100 504 883
www.petermac.org

Locations
East Melbourne
Bendigo
Box Hill
Moorabbin
Sunshine



30th March 2014

To whom it may concern,

Re: James Hadfield
Letter of Support
Contribution as co-author of Murtaza et. al. Nature 2013

Murtaza et. al. "Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA" was published in Nature in 2013. The manuscript, for which I was a co-first author, demonstrated the use of exome sequencing of circulating tumour DNA (ctDNA) to study mutational evolution in patients receiving treatment for advanced solid malignancies. This landmark study revealed the potential of ctDNA to provide a 'liquid biopsy' alternative to tissue biopsies for the monitoring of tumour evolution.

Circulating DNA is highly fragmented and is usually present at levels of only a few thousand amplifiable copies per ml of plasma. Importantly, only a small fraction of this DNA may be derived from the tumour, amongst a large pool of wild-type DNA released from healthy cells. Genome wide approaches to detect somatic mutations in plasma have traditionally been challenging to perform due to the limited and variable amount of ctDNA present in a given plasma sample. A further challenge in performing exome sequencing of plasma DNA is the initial amount of DNA material required. Typically, one ml of plasma in a healthy individual contains around 3000 copies of the genome (i.e. ~10ng/ml). Although cancer patients may have 10 times higher levels of circulating DNA this is still much smaller than the amount of DNA usually required for standard next generation sequencing library preparation protocols.

In this context, James was instrumental in the development of a successful method to allow exome sequencing of low amounts of plasma DNA. His expertise in genomic technologies led us to the ThruPLEX (Rubicon Genomics) protocol that allows library preparation using small amounts of DNA. With his close involvement we were able to demonstrate the application of the ThruPLEX protocol to perform exome sequencing of plasma DNA from as little as 2ng DNA. James oversaw aspects of the library preparation and sequencing for this study, and provided important contributions during the preparation of the manuscript.

In this study, the genome-wide analysis of plasma DNA allowed us to study changes in mutational profiles during the development of acquired treatment resistance in patients with advanced solid malignancies. This work has established the role of ctDNA analysis as a unique tool to study clonal evolution during disease progression and treatment. Overall, James has made a critical contribution to this work, which now provides the foundation for future research using ctDNA to address fundamental biological and clinical questions.

Peter MacCallum Cancer Centre

St Andrews Place
East Melbourne Victoria
Postal Address
Locked Bag 1 A'Beckett Street
Victoria 8006 Australia
Phone +61 3 9656 1111
Fax +61 3 9656 1400
ABN 42 100 504 883

www.petermac.org

Locations

East Melbourne
Bendigo
Box Hill
Moorabbin
Sunshine



Yours sincerely,

Dr Sarah-Jane Dawson (MBBS, FRACP, PhD)
Consultant Oncologist &
Group Leader, Molecular Biomarkers and Translational Genomics Laboratory
Peter MacCallum Cancer Centre
St Andrew's Place, East Melbourne
Victoria 3002, AUSTRALIA
Phone +61 3 9656 3728
E-mail Sarah-Jane.Dawson@petermac.org



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre
Robinson Way
Cambridge
CB2 0RE

28 March 2014

Dear Sir/Madam,

I am writing with reference to the following publication:

M. Murtaza, S.-J. Dawson, D. W. Y. Tsui, D. Gale, T. Forshew, A. M. Piskorz, C. Parkinson, S.-F. Chin, Z. Kingsbury, A. S.C. Wong, F. Marass, S. Humphray, J. Hadfield, D. Bentley, T. M. Chin, J. D. Brenton, C. Caldas, N. Rosenfeld. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 2013 497 (7447), 108-112

At the time this paper were written, James Hadfield and I were working at the Cancer Research Cambridge Institute, University of Cambridge (Formerly known as Cancer Research UK Cambridge Research Institute). At that time, James were the Head of the Genomic Core facility of the abovementioned institute.

The research describes a novel method to perform exome-wide analysis of cancer genome by simply testing the blood plasma instead of tumour biopsies. It opens up opportunities to perform non-invasive "liquid biopsies" at disease relapse to identify the mechanisms behind resistance to cancer drug and thereby facilitating the development of tailor-made personalised medicine.

James' major contributions to the research were in designing the procedures for DNA library preparation and exome capture analysis, with particular emphasis in negotiating with appropriate companies to get early access to the chemistry which allows most efficient processing and preparation of plasma materials for downstream sequencing analysis. James also assisted in writing the manuscript and approved the final version.

Yours

Dana WY Tsui

Postdoctoral Research Fellow
Cancer Research UK Cambridge Institute

Dana.tsui@cruk.cam.ac.uk
+44(0)1223769736

8 January 2015

To whom it may concern,

Re: Contribution of James Hadfield to publication by
Murtaza et al., Nature 2013

Dear Sir or Madam:

In reference to the following publication, of which I am a corresponding and senior author:

M. Murtaza, S.-J. Dawson, D. W. Y. Tsui, D. Gale, T. Forshaw, A. M. Piskorz, C. Parkinson, S.-F. Chin, Z. Kingsbury, A. S.C. Wong, F. Marass, S. Humphray, J. Hadfield, D. Bentley, T. M. Chin, J. D. Brenton, C. Caldas, N. Rosenfeld. Non- invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 2013 497 (7447), 108-112

I would like to write to highlight the contributions of James Hadfield.

At the time this work was developed and performed and the paper was written, I was a Junior Group Leader at the Cambridge Research Institute (now the Cancer Research UK Cambridge Institute), where James Hadfield was head of the Genomics core facility.

This work was based on application of novel library preparation and sequencing methodologies to well-characterised cohorts of samples of circulating plasma DNA from cancer patients. James Hadfield played important roles in identifying, selecting and coordinating use of appropriate technologies, which was critical to the success of this project, since routinely-used technologies at the time were not able to provide the required data by sequencing circulating DNA. James contributed to experiments and data analysis, reviewing and approving the manuscript.

This paper, published in 2013, is highly influential in the development of the circulating tumour DNA field. As evidence of its impact, it was reviewed and covered by news items in high-circulating newspapers such as The Times and high impact professional media such as Nature Medicine and GenomeWeb. In the short time that has elapsed it has already been cited more than 170 times in the professional literature. My research lab has relied on the approaches described in the paper to secure significant further highly-competitive funding, to expand this work and apply this to additional cancer types.

Sincerely,



Nitzan Rosenfeld

Senior Group Leader

Cancer Research UK Cambridge Institute, University of Cambridge

Tel: 01223-769769, Fax: 01223-769510, e-mail: nitzan.rosenfeld@cruk.cam.ac.uk

Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way
Cambridge CB2 0RE

Tel: +44 (0)1223 769 500

www.cruk.cam.ac.uk

11 September 2014

Re: Idris et al, 1996

Dear Sir/Madam,

I am writing with reference to the following scientific paper ...

Idris SF, Ahmad SS, Scott MA, Vassiliou GS, Hadfield J. The role of high-throughput technologies in clinical cancer genomics. Expert Rev Mol Diagn. 2013 Mar;13(2):167-81

As co-author of the above paper, this letter is to define the critical role Dr James Hadfield had in it's conception and execution. James was the lead author of this paper and initially approached all of the above co-authors regarding the idea. He worked to coordinate everyones individual roles within writing the manuscript and designed a number of excellent, original figures which were used within it. His expertise on genomic sequencing, both in terms of current and future technologies was instrumental in producing the final article. He was responsible for more than half of the final text and reviewed the whole paper as final author.

Throughout my time working with James I was very much impressed with his knowledge and ability to communicate complex information.

Yours sincerely,



Dr Saif Ahmad
Spr in Clinical Oncology and Cambridge Cancer Centre PhD Clinical Fellow

University of East Anglia
Postgraduate Research Service
Norwich Research Park
NORWICH
NR4 7TJ UK

20th March 2014

Dear Sirs/Madams,

Re: James Hadfield – PhD by Publications, Expert Review Paper
The role of High-Throughput Technologies in clinical cancer genomics
Idris SF, Ahmad SS, Scott MA, Vassiliou GS, Hadfield J
Expert Rev Mol Diagn. 2013 Mar;13(2):167-81.

I am writing with reference to the above scientific paper which I co-authored with Mr James Hadfield and others. The paper was an expert review which Mr Hadfield was invited to author by a member of Editorial team for *Expert Review of Molecular Diagnostics*, in recognition of his leading contributions to the field. I was invited by him to be a co-author because of my expertise in the pathogenesis and diagnosis of haematological cancers.

As co-author of the above paper, I can confirm that Mr Hadfield, as senior author, was the driving force behind the paper. He defined the papers structure, personally wrote significant parts of the paper and guided the remaining authors through what is a very large and rapidly changing field of study. Mr Hadfield's own expertise in the applications of high-throughput technologies was key for putting together this expert review which has been well received in the field and already has had a number of citations despite its recent publication.

Yours Sincerely

Dr George Vassiliou

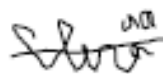
Dear Sir or Madam,

I am writing this letter as the lead co-author of Azizan et al: Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension; Nature Genetics, 2013, to confirm Mr James Hadfield's contribution to the work. Myself and Prof Morris Brown approach Mr Hadfield after reviews of our manuscript suggested a need for validation of results through sequencing. At the initial meeting with James he explained the multiple options we might consider and suggested we use the Fluidigm micro-fluidic system. He also arranged for us to collaborate with Dr Nitzan Rosenfeld group on automated design of PCR amplicons. James was involved in the design and interpretation of the microfluidic sequencing for our paper, where we found an additional 3 novel mutations in ATP1A1 or CACNA1D. He personally performed the actual Access Array processing, barcode-indexing and Illumina sequencing of the samples included in the paper. James also helped in the preparation of the microfluidic sequencing methods section of the paper. Without James's help and experience we would have had a significantly harder time in executing this validation.

The microfluidic sequencing was a complement to Exome sequencing and TaqMan genotyping and helped to demonstrate that we had found all possible mutations in the gene of interest to our study.

James was really helpful and working with him was good fun.

Yours Sincerely,



Dr. Elena Aisha Binti Azizan

Elena.azizan@gmail.com

Morris J Brown MA MSc MD FRCP FAHA FMedSci

Professor of Clinical Pharmacology

Anthony P Davenport MA PhD DIC FBPharmacoS

Reader in Cardiovascular Pharmacology

Kevin M O'Shaughnessy MA DPhil MRCP

Senior Lecturer/Honorary Consultant Physician

01223 762577 mjb14@medschl.cam.ac.uk

01223 336899 apd10@medschl.cam.ac.uk

01223 762578 kmo22@cam.ac.uk



**UNIVERSITY OF
CAMBRIDGE**

**Clinical Pharmacology Unit
Department of Medicine**

16th June 2014

Dear Sir or Madam,

I am writing this letter as the corresponding author of Azizan et al: Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension; Nature Genetics, 2013, and am happy to confirm James's contribution to the final published work. I initially approach Mr Hadfield to discuss methods for high-throughput validation and screening of three genes ATP1A1, CACNA1D and KCNJ5 in a patient cohort. James explained the different options we might consider and suggested we use the Fluidigm Access Array system and Illumina sequencing for targeted PCR-amplicon sequencing. He very helpfully put us in touch directly with Dr Nitzan Rosenfeld's group who we collaborated with on the automated design of PCR amplicons. James performed the majority of Access Array processing, barcode-indexing and Illumina sequencing of the samples included in the paper. James also helped the lead-author of the paper with the materials and methods section of the paper. James's significant expertise in next-generation sequencing made our validation experiments much easier than might otherwise have been the case. I would estimate he contributed to 50% of the microfluidic sequencing validation, and although that was only a very small part of the overall publication it was an important contribution.

I am happy to write this letter of support for James's PhD by Publication. Please do contact me directly if you have any further questions.

Yours sincerely

Morris Brown

12 January 2015

University of East Anglia
Norwich Research Park
Norwich
NR4 7TJ

Dear Sir/Madam,

I am writing in support of James Hadfield's PhD by publication and in particular with regard to the following publication that James and I co-authored:

James Hadfield and Matthew D. Eldridge, **Multi-genome alignment for quality control and contamination screening of next-generation sequencing data**, *Front. Genet.* 5:31 (2014)

James originally conceived the idea of a contamination screen that could be used routinely as a data quality control step for the sequencing data his Genomics Core facility produces, following a series of incidents that in which the sequencing data originating from one laboratory were shown to have significant levels of unintended sequence from another species. We worked together on designing a bioinformatic analysis approach that sampled and aligned sequence reads against a panel of reference genomes including all the species regularly sequenced in our institute as well as all those available for bacterial, viral and fungal species. This developed into a more sophisticated tool that also identified the amount of adapter sequence present in the sequence reads.

James played a significant role in helping to design the visualization that summarizes the series of alignment results for each dataset in a single figure. The end result is a single figure that captures most of the key metrics of a sequencing run including the numbers of sequence reads, the error or mismatch rates indicative of sequence data quality, the relative proportions of reads aligning to expected and contaminant species, and the amount of adapter sequence content.

James led the preparation of the manuscript and was responsible for the background and context sections as well as the discussion of the results and how the tool has been and could be applied. I contributed the methods section and prepared the screenshots for some representative sequencing runs/datasets.

The multi-genome alignment (MGA) tool has been used at the Cancer Research UK Cambridge Institute for all sequencing runs over the past 5 years, within an automated post-processing pipeline. It is one of the primary outputs used by James' facility for assessing data quality and is the result of a highly cooperative and productive relationship between the Genomics and Bioinformatics Core facilities that James and I manage respectively.

The publication received considerable attention from the genomic sequencing community and ranks among the top 2% of all articles published by Frontiers in Genetics based on Altmetric score, a measure of the quality and quantity of online attention articles receive.

Yours faithfully,



Dr Matthew Eldridge
Head of Bioinformatics, CRUK Cambridge Institute

Appendix 2: Publications submitted

Copies of all published work are reproduced here with permission of the journals.

A differential PCR assay for the detection of c-erbB 2 amplification used in a prospective study of breast cancer

B A Jennings, J E Hadfield, S D Worsley, A Girling, G Willis

Abstract

Aims—To establish a robust differential polymerase chain reaction (PCR) assay for the detection of c-erbB 2 amplification in breast cancer that can be used in a routine pathology laboratory. Once established, the assay was used in a prospective study of breast tumours to investigate the relation between c-erbB 2 amplification and both recognised prognostic features and short term clinical outcome.

Methods—The differential PCR was used for the co-amplification of c-erbB 2 and a reference gene from 48 tumour DNA samples and control DNA samples. The ratio of the two genes was determined by image analysis of the PCR products electrophoresed on a highly resolving agarose gel.

Results—The differential PCR assay was shown to be accurate and reproducible using the conditions outlined. Twenty six per cent of the breast cancer patients were shown to have c-erbB 2 amplification in their tumour biopsies. Twenty eight per cent of the patients died of their disease or had disease recurrence during the follow up period and 73% of these patients had amplification of c-erbB 2.

Conclusions—A significant association was found between c-erbB 2 amplification and early disease recurrence. This assay could be used to provide a marker for poor prognosis in breast cancer.

(*J Clin Pathol: Mol Pathol* 1997;50:254-256)

Keywords: differential PCR; c-erbB 2; breast cancer

two PCR product bands visualised on a gel. Other studies have examined a variety of differential PCR methods⁵⁻¹¹ for the detection of c-erbB 2 amplification, which has been shown to correlate with p185^{c-erbB2} immunostaining,^{8,9} but only a few studies have examined the prognostic use of the assay with clinical follow up.^{5,9,11}

In this study, reliable measurement of c-erbB 2 amplification was achieved when the co-amplification of the two gene sequences (c-erbB 2 and β globin) was optimised. The gene targets in this study are on different chromosomes and so the results will reflect an increase in the c-erbB 2 copy number irrespective of whether a small region of the chromosome or the whole of chromosome 17 is duplicated. Chromosome aneuploidy, including loss and gain of chromosome 17, is seen frequently in breast cancer.

The two primer pairs were selected to be non-complementary at their 3' termini and for their similar GC content. In addition, the PCR amplification was stopped before the end of the exponential phase of the reaction (the plateau) was reached.

We have analysed DNA extracted from 42 breast tumour samples, most of which have been described previously.^{12,13} Short term clinical follow up of the breast cancer patients revealed that c-erbB 2 amplification was associated strongly with early relapse. These data demonstrate that this assay identifies a subset of breast cancer patients with poor short term prognosis.

Methods

Samples from 42 female patients treated for primary breast cancer by the same surgical team between 1993 and 1994 were included in this study. The mean age of the patients was 63 years, ranging from 35 to 85 years. No woman had received preoperative radiotherapy. Fresh tumour samples were obtained from both mastectomy and excision biopsy specimens and DNA was extracted as described previously.¹² In addition to routine pathology examination, samples used for DNA extraction were examined histologically and were shown to consist of at least 70% tumour cells. The tumours ranged in size from 0.7 cm to 10 cm (mean 3.23). The selection criterion for inclusion in the study was that an adequate amount of tumour was available for the extraction of DNA.

DNA was also extracted from the peripheral blood of 10 of the breast cancer patients and 28

The c-erbB 2 oncogene, also known as HER2 and neu, is located on chromosome 17 (q21-22) and encodes a 185 kDa transmembrane protein that is a member of the erbB family of receptor tyrosine kinases.^{1,2} Frequently, c-erbB 2 is amplified and overexpressed in breast cancer and both of these abnormalities have been found to correlate with both disease recurrence and reduced overall survival.³⁻⁶

This paper describes a simple, robust, and highly sensitive differential polymerase chain reaction (PCR) method for detecting amplification of c-erbB 2 in a routine pathology laboratory. Differential PCR is a semiquantitative assay for the co-amplification of a target gene and a reference gene in the same reaction tube.⁷ The level of amplification of the target gene is seen by the ratio between the intensity of the

**Molecular Genetics
Department, Norfolk
and Norwich Hospital,
Norwich, UK**

B A Jennings
J E Hadfield
S D Worsley
G Willis

**Histopathology
Department**
A Girling

Correspondence to:
Dr Jennings, Molecular
Genetics Department,
Norfolk and Norwich
Hospital, Brunswick Road,
Norwich NR1 3SR, UK.

Accepted for publication
8 July 1997

normal individuals. In addition, DNA was extracted from two cell lines with known alterations of *c-erbB 2* copy numbers: MCF7 may be hemizygous for *c-erbB2* and SKBR3 has up to eightfold *c-erbB 2* amplification.⁷ These samples served as controls for the optimisation of the assay.

All DNA was diluted to the same concentration (50 ng/μl). A dilution series of the SKBR3 DNA in normal DNA was prepared to test the linearity of the measurement achieved by the differential PCR.

The primers used were as follows. For *c-erbB 2*: 5'-TCGGAACGTGCTGGTCAAGA-3' (sense primer) and 5'-ATGGTACTCTGTCTCGTCAA-3' (antisense primer); these primers amplify a 91 base pair fragment from exon 3.

For β globin: 5'-ACACAACCTGTGTTCACTAGC-3' (sense primer) and 5'-CAACTTCATCCACGTTCCACC-3' (antisense primer); these primers amplify a 110 base pair fragment from exon 1.

DNA amplification was carried out in duplicate for each sample using a Progene thermal cycler (Technique, Cambridge, UK). Each 50 μl reaction mixture contained 25 μl PCR master mix (Boehringer Mannheim, Lewes, East Sussex, UK), 5 μl of each primer (50 pmol), 2 μl of DNA (100 ng), and 13 μl of sterile distilled water. Two controls that contained all the reagents but no target DNA were included with each batch. The reaction mixtures were prepared and kept on ice until the heating block of the thermal cycler reached the denaturation temperature (94°C). Each reaction mixture was placed at 94°C for five minutes and then subjected to 35 amplification cycles; each cycle was 30 seconds at 94°C, 30 seconds at 50°C, and 30 seconds at 72°C. This was followed by a final extension at 72°C for seven minutes. Initially, the optimum number of PCR cycles was determined empirically by analysing the amplification products after 20 to 50 cycles, at five cycle increments.

Amplification products were separated by electrophoresis using a 3% metaphor agarose gel (Flowgen, Lichfield, Staffordshire, UK), stained with SYBR green DNA gel stain (Flowgen), and visualised by ultraviolet illumination. The sizes of the PCR products were compared with a molecular weight marker, pUC18 DNA digested with HaeIII (Sigma, Poole, Dorset, UK). The gel images were captured using a CCD camera linked to an image processing system (GDS 8000; UVP, Cambridge, UK). The intensity of the *c-erbB 2* band and the β globin band was determined for each specimen, by means of Gelworks software (UVP). These results were expressed as the ratio: intensity of the *c-erbB 2* band/intensity of the β globin band. The ratios determined for the tumour samples were converted into a measure of gene amplification using the ratios determined for the normal controls and cell line controls. The cut off point for amplification was the mean of the normal range plus two standard deviations (SD).

Univariate statistical analysis comparing clinical and laboratory findings was carried out using Fisher's exact test.

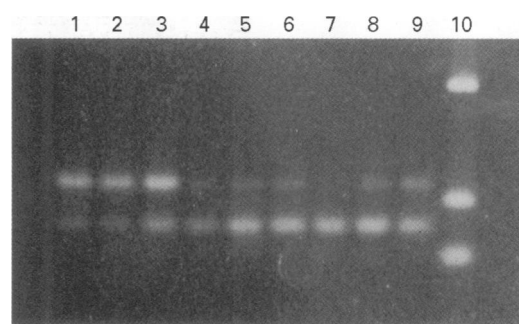


Figure 1 The amplification products from the differential PCR of *c-erbB 2* and the reference gene β globin from three normal DNA controls (lanes 1–3), from three breast tumours with amplification of *c-erbB 2* (lanes 4–6), and from a dilution of SKBR3 DNA in normal DNA, equivalent to eight copies (lane 7), five copies (lane 8), and three copies (lane 9) of *c-erbB 2*. The 174, 102, and 80 base pair bands from the molecular weight marker (pUC18 DNA digested with HaeIII) are seen in lane 10.

Results

The results relating to typical control DNA and breast tumour DNA samples are shown in fig 1. The 91 and 110 base pair PCR products were resolved clearly using 3% metaphor agarose. For all normal DNA controls the intensity of the higher molecular weight β globin band was greater than the intensity of the lower molecular weight *c-erbB 2* band.

The study of the optimum numbers of PCR cycles showed that consistent differences in the *c-erbB 2* and β globin PCR products could be observed from 25 cycles (the sensitivity limit of the assay) to 40 cycles (the PCR plateau).

A dilution series consisting of five twofold dilutions of SKBR3 DNA into normal DNA was used for differential PCR. The ratios of *c-erbB 2* to β globin showed a linear relation with the number of copies of *c-erbB 2* present (examples are shown in fig 1). This demonstrated the quantitative accuracy of this differential PCR method. Each sample was analysed at least twice and each replicate gave concordant results.

Forty two DNA samples from primary tumours and six DNA samples from nodal metastases were analysed for *c-erbB 2* amplification. The distribution of *c-erbB 2*/ β globin ratios was bimodal. The majority of samples had ratios similar to the normal controls (within the normal range of mean +2 SD) and the remainder had from three to greater than eightfold *c-erbB 2* amplification. Eleven primary tumours and lymph node metastases derived from two of these tumours had *c-erbB 2* amplification. Thirty one primary tumours and lymph node metastases derived from four of these tumours had a normal *c-erbB 2* copy number. Therefore, 11 of 42 (26%) of these breast cancer patients had *c-erbB 2* amplification in their tumours. These results are shown in relation to the tumour types in table 1.

Clinical follow up information was available for 41 of 42 patients. The median duration of follow up was 28 months with a minimum of six months for a patient who died and a maximum of 42 months. Seven patients died of breast cancer and a further four had recurrent disease. One patient died of other causes and so was excluded from statistical analysis.

Table 1 Summary of the tumours with c-erbB 2 amplification in relation to tumour type and grade

Tumour type	Total number of tumours	Tumours with c-erbB 2 amplification
Ductal grade 1	6	1
Ductal grade 2	10	2
Ductal grade 3	19	6
Lobular	5	1
Special type	1	0
Other ¹³	1	1

Eleven of 40 (28%) patients died of their disease or had disease recurrence and 8 of 11 (73%) of these patients had c-erbB 2 amplification. The latter was associated strongly with early disease recurrence ($p = 0.0003$).

Twenty three patients had histological evidence of lymph node metastases at presentation and eight of these patients (35%) died of their disease or had disease recurrence. Lymph node metastasis had an association with early disease recurrence but this did not reach statistical significance ($p = 0.1$). Six of seven (86%) patients with c-erbB 2 amplification and lymph node metastases died of their disease or had disease recurrence.

Discussion

We present a robust differential PCR assay for the detection of c-erbB 2 amplification in human DNA. Because a numerical result is generated by an image analysis system, this technique provides objective analysis of a molecular marker, making it a good candidate for development as a routine pathology test. The PCR primers and cycle numbers have been optimised to give reproducible results that are not subject to primer dimer or plateau effect artefacts. The sensitivity of this assay, with the use of SYBR green DNA stain and the small sizes of the PCR products generated by the chosen primers make the protocol amenable to the analysis of small amounts of highly degraded DNA, as described by others.^{8 14}

We found that the concentration of DNA must be standardised for each batch of samples analysed. This may be necessary to avoid differential chelation of magnesium ions by DNA and to ensure that all reactions remain within the exponential phase of the reaction. The importance of using a standard DNA template concentration for accurate differential PCR has also been shown in other studies.¹⁴

c-erbB 2 amplification was detected in 11 of 42 (26%) of the primary breast tumours, a similar finding to other studies.^{9 11} Eight sets of primary breast carcinoma and their nodal metastases were analysed for gene amplification. There was no evidence for an alteration in gene copy number between the primary and secondary tumour. This suggests that any alteration to c-erbB 2 occurred before and was maintained during metastasis. No significant correlation was found between c-erbB 2 amplification and the standard histopathological prognostic markers: tumour size, type, grade, lymph node status, and oestrogen receptor status.

After a short clinical follow up, this study indicates that c-erbB 2 amplification is associated with more aggressive tumours because

gene amplification was associated significantly with a poor prognosis, assessed by death and/or disease recurrence ($p = 0.0003$). This finding concurs with those of some,^{5 11} but not all,⁹ other studies that have used differential PCR methods in retrospective analyses with longer periods of clinical follow up than is presented here. In this study, the DNA analysed was from symptomatic patients, many of whom presented with relatively large and intermediate or high grade tumours. It would also be interesting to investigate the prognostic use of this assay in the often smaller and better differentiated tumours detected in the national breast screening programme.

Breast cancer is a heterogeneous disease. There may be many different mechanisms by which tumours grow, metastasise, and evade treatment response. Genetic markers that subclassify these tumours could help to identify those patients who would benefit most from adjuvant therapy.

We thank the Big C Appeal for funding this project and Mr David Ralphs and his team from the Norfolk and Norwich Hospital for providing the surgical biopsies. We are also grateful to Dr Samir Alhasan, Wayne State University, Detroit, USA and the CRC Human Cancer Genetics Research Group, Addenbrooke's Hospital, Cambridge for the gifts of the cell line DNA samples.

- Coussens L, Yang-Feng TL, Liao YC, Chen E, Gray A, McGrath J, *et al.* Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science* 1985;230:1132-9.
- Schechter AL, Stern DF, Vaidyanathan L, Decker SL, Drebin JA, Greene MI, *et al.* The neu oncogene: an erb-B-related gene encoding a 185,000-Mr tumour antigen. *Nature* 1984;312:513-16.
- Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987;235:177-82.
- Gullick WJ, Love SB, Wright C, Barnes DM, Gusterson B, Harris AL, *et al.* c-erbB2 protein overexpression in breast cancer is a risk factor in patients with involved and uninvolved lymph nodes. *Br J Cancer* 1991;63:434-8.
- Lonn U, Lonn S, Nilsson B, Silversward C, Stenkvist B. Demonstration of gene-amplification by PCR in archival paraffin-embedded breast cancer tissue. *Breast Cancer Res Treat* 1994;30:147-52.
- Molland JG, Barraclough BH, Gebiski V, Milliken J, Bilous M. Prognostic significance of c-erbB2 oncogene in axillary node-negative breast cancer. *Aust NZ J Surg* 1996;66:64-70.
- Frye RA, Benz CC, Liu E. Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene* 1989;4:1153-7.
- Gramlich TL, Cohen C, Fritsch C, DeRose PB, Gansler E. Evaluation of c-erbB2 amplification in breast carcinoma by differential polymerase chain reaction. *Am J Clin Pathol* 1994;101:493-9.
- An H, Niederacher D, Beckmann MW, Gohring UJ, Scharl A, Picard F, *et al.* ERBB2 gene amplification detected by fluorescent differential polymerase chain reaction in paraffin-embedded breast carcinoma tissues. *Int J Cancer* 1995;64:291-7.
- Friedrichs K, Lohmann D, Hoffer H. Detection of HER-2 oncogene amplification in breast cancer by differential polymerase chain reaction from single cryosections. *Virchows Archiv B Cell Pathol* 1993;64:209-12.
- Lonn U, Lonn S, Nilsson B, Stenkvist B. Prognostic value of erb-B2 and myc amplification in breast cancer imprints. *Cancer* 1995;75:2681-7.
- Worsley SD, Jennings BA, Khalil KH, Mole M, Girling AC. Cyclin D1 amplification and expression in human breast carcinoma: correlation with histological prognostic markers and oestrogen receptor expression. *J Clin Pathol: Mol Pathol* 1996;49:M46-50.
- Killick SB, McCann BG. Osteosarcoma of the breast associated with fibroadenoma. *Clin Oncol* 1995;7:132-3.
- Neubauer A, Neubauer B, He M, Effert P, Iglehart D, Frye RA, *et al.* Analysis of gene amplification in archival tissue by differential polymerase chain reaction. *Oncogene* 1992;7:1019-25.



ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions

Dominic Schmidt^{a,b}, Michael D. Wilson^b, Christiana Spyrou^{b,c}, Gordon D. Brown^b,
James Hadfield^b, Duncan T. Odom^{a,b,*}

^a Department of Oncology, Hutchison/MRC Research Centre, Hills Road, Cambridge, CB2 0XZ, UK

^b Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

^c Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB3 0WY, UK

ARTICLE INFO

Article history:

Accepted 1 March 2009

Available online 9 March 2009

Keywords:

ChIPseq

ChIP-seq

Chromatin immunoprecipitation

High-throughput sequencing

ABSTRACT

Chromatin immunoprecipitation (ChIP) allows specific protein–DNA interactions to be isolated. Combining ChIP with high-throughput sequencing reveals the DNA sequence involved in these interactions. Here, we describe how to perform ChIP-seq starting with whole tissues or cell lines, and ending with millions of short sequencing tags that can be aligned to the reference genome of the species under investigation. We also outline additional procedures to recover ChIP-chip libraries for ChIP-seq and discuss contemporary issues in data analysis.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Protein–DNA interactions play vital roles in the regulation of gene expression, genome integrity and chromatin organization. The *in vivo* mapping of transcription factor binding and modified histones has greatly broadened our understanding of how the genome can be deployed to achieve tissue and developmental stage-specific gene regulation. Computational methods have provided substantial insight into our understanding of transcriptional regulation [1], and yet recent experimental discoveries have underscored the need for a simple and reproducible method for mapping protein–DNA interactions on a global basis. These include recent discoveries that: (i) sequence-specific transcription factors (TFs) do not occupy all positions in the genome that would be predicted by their corresponding binding matrices [2,3], (ii) sequence specific transcription factors often bind regions that do not show similarity to their canonical binding matrices [2–5] and (iii) the binding patterns of TFs between species are poorly conserved [6–8].

Chromatin Immunoprecipitation (ChIP) [9,10] is a commonly used technique to detect interactions between proteins and DNA, which is based on the enrichment of DNA associated with a protein of interest. The development of ChIP coupled with high-throughput sequencing analysis (ChIP-seq) allows the unbiased identification of binding sites of a given transcription factor and has

overcome several limitations inherent to microarray analysis of ChIP (ChIP-chip) [11,12].

Due to their size and more repetitive nature, higher eukaryotic genomes are a challenge for tiling microarray design. Most of the repetitive sequence cannot be interrogated with high confidence, whereas direct sequencing can reveal binding events located in repetitive regions in the mammalian genome [13–15]. Every model organism requires species-specific microarray designs before ChIP-chip can be performed, while ChIP-seq can be done without prior knowledge of the underlying sequence and relies only on the subsequent DNA sequence alignment to the reference genome of interest. Furthermore, the nature of the microarray hybridization signal makes detection and rigorous quantification of low abundance signals problematic. Taken together, ChIP-seq can provide greater resolution, sensitivity and specificity compared to ChIP-chip [11,14,16].

A number of high-throughput sequencing technology platforms have been developed that are suitable for ChIP-seq, including the Genome Analyzer (Illumina, formerly Solexa), SOLiD (Applied Biosystems), 454-FLX (Roche) and HeliScope (Helicos) [17]. The Illumina Genome Analyzer and the ABI SOLiD sequencers produce shorter reads but give a higher number of sequencing reads per run, whereas the 454-FLX sequencer gives longer yet fewer sequencing reads per run [18]. Sequencing depth is a critical factor in identifying weaker binding positions and it has been shown that millions of mapped sequencing tags are needed to detect enrichments significantly higher than twofold [19].

Here, we outline detailed methodologies for ChIP-seq using the Illumina Genome Analyzer to produce tens of millions of aligned sequencing tags. Our protocol adapts methods described

* Corresponding author. Address: Department of Oncology, Hutchison/MRC Research Centre, Hills Road, Cambridge, CB2 0XZ, UK. Fax: +44 0 1223 404199.

E-mail address: duncan.odom@cancer.org.uk (D.T. Odom).

previously [14,20] with additional modifications and technical improvements to the chromatin immunoprecipitation (ChIP) and library generation steps.

2. Description of method

2.1. Overview

A successful ChIP experiment begins with the crosslinking of protein–DNA interactions using formaldehyde (Fig. 1). Histone modifications can also be successfully identified using non-crosslinked native chromatin in the ChIP protocol [21], but the ability to capture weaker and transient protein–DNA interactions has made formaldehyde fixation of starting materials a standard practice. After crosslinking, the tissue is homogenized, and the cells are lysed. Subsequently, the chromatin is sheared using sonication and incubated with magnetic beads coupled to an antibody specific for the target protein. The success of the ChIP is dependent on the antibody being used; indeed, we have found that a large fraction of highly specific, IP-proven antisera do not perform well against

crosslinked chromatin. We therefore strongly recommend the use of a positive control antibody as described below when testing new antibodies, testing collected tissues, or performing ChIP-seq for the first time. In principle, the generation of a sequencing library from DNA is relatively straightforward. However, as opposed to ChIP analyzed by real-time PCR, ChIP-seq requires a larger quantity of precipitated DNA to minimize the generation of adapter dimer artefacts and to preserve the complexity of the DNA sample. This protocol is routinely used in our laboratory and has been successful with a variety of antibodies, tissues and cells from a wide range of vertebrate species.

2.2. Step-by-step protocol

2.2.1. Crosslinking of cells or primary tissues

Covalent fixation of the protein–DNA complexes is achieved by brief formaldehyde fixation. Ideally the starting material for one ChIP uses 5×10^7 cells from culture or the equivalent of one-quarter of an adult mouse liver. While it is possible to start with limited material [22,23], we have found that higher amounts of starting

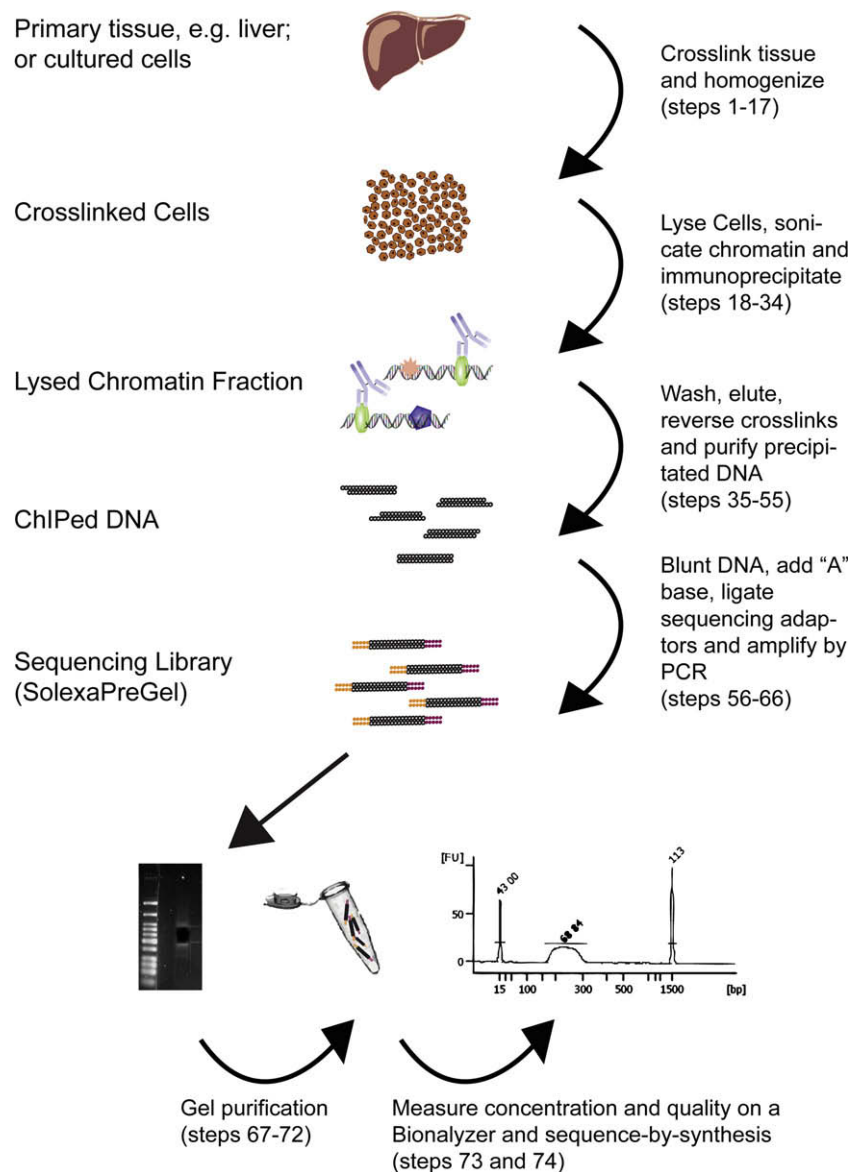


Fig. 1. Outline of ChIP-seq procedure.

material yield more consistent and reproducible protein–DNA enrichments. To crosslink material for ChIP, follow steps 1–6 for cultured cells and steps 7–17 for whole tissue.

2.2.1.1. Cells

1. Add 1/10 volume of fresh 11% formaldehyde solution (50 mM Hepes–KOH, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 11% formaldehyde) to plates or flasks. Alternatively, pour off cell culture media and cover cells in a solution of 1% formaldehyde (final concentration) in 50 mM Hepes–KOH, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA.
2. Swirl briefly and let sit at room temperature for 10 min.
3. Add 1/20 volume of 2.5 M glycine to quench formaldehyde.
4. Rinse cells twice with ice cold PBS.
5. Transfer cells to 15 ml conical tubes and spin 4 min at $2000 \times \text{rcf}$.
6. Proceed with cell lysis or freeze cells in liquid nitrogen and store pellets at -80°C . Continue with step 18.

2.2.1.2. Primary tissue

7. Whenever possible, perfuse tissue with PBS to remove blood.
8. On a kimwipe wetted with PBS, mince tissue quickly with a razorblade into small pieces. The pieces should be not bigger than 0.5 cm^3 .
9. Add tissue to at least five volumes of freshly prepared solution A (1% formaldehyde, 50 mM Hepes–KOH, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA).
10. Mix and leave at room temperature for 20 min.
11. Add 1/20 volume of 2.5 M glycine to quench formaldehyde.
12. Rinse tissue with PBS and flash freeze or proceed directly to step 13.
13. Dounce tissue in ice-cold PBS first with the loose and later with the tight pestle (Dounce Tissue Grinder from Wheaton Science, Catalog #357544). We do not add protease inhibitors during this step.
14. The equivalent of one dounced mouse liver is filtered into a 50 ml conical tube through a $100 \mu\text{m}$ cell strainer to remove connective tissue. Fill tube with ice-cold PBS to 40 ml and centrifuge at 4°C at $2500 \times \text{rcf}$ for 3 min.
15. Discard supernatant and repeat wash.
16. Resuspend pellet in 10 ml ice cold PBS and transfer to 15 ml conical tube, centrifuge as above. Distribute into several 15 ml tubes if there would be more than 2 ml of tissue per tube.
17. Proceed with lysis, or freeze cells in liquid nitrogen and store pellets at -80°C .

2.2.2. Preblock and binding of antibody to magnetic beads

Like all immunoprecipitation experiments, successful ChIP requires a suitable antibody. With ambitious antibody generation efforts led by both academic and industrial labs, many candidate antibodies corresponding to DNA binding proteins are available. Numerous antibodies have been shown to work in ChIP; nevertheless, it is often the case that a series of antibodies must be tested against a protein of interest. Often the creation of new antisera targeted to different epitopes is required to create ChIP-grade antibodies. When testing new antibodies or performing ChIP (and especially ChIP–seq) for the first time we recommend using a positive control such as anti-H3K4me3 (ab8580, Abcam) which detects the tri-methylated lysine 4 form of histone H3 in a wide range of species, provides robust reliable enrichments, and highlights potential transcription start sites in the genome.

Magnetic beads are less porous than traditional agarose beads [22,21] and easier to handle, and hence highly recommended for chromatin immunoprecipitation to reduce background precipita-

tion of nonspecific DNA. The exact type of magnetic beads depends of the species and subclass of the antibody being used. Protein G coated beads have high affinity to most rabbit and goat antibodies. Antibodies are incubated with the magnetic beads prior to the addition of the nuclear extracts, and excess, unbound antibodies are then washed away. This ensures that unbound antibodies cannot compete with the antibodies attached to the magnetic beads for target epitopes during ChIP. All antibody incubations and washes are performed at 4°C .

18. Add 100 μl magnetic beads (Invitrogen, Dynabeads) to a 1.5 ml microfuge tube. Add 1 ml block solution (0.5% BSA (w/v) in PBS). Set up 1 tube per IP.
19. Collect the beads using magnetic stand. Remove supernatant by aspiration.
20. Wash beads in 1.0 ml block solution two more times.
21. Resuspend beads in block solution and add 2–15 μg of antibody in a final volume of 250 μl .
22. Incubate overnight or a minimum of 4 h on a rotating platform at 4°C .
23. Wash magnetic beads as described above (3 times in 1 ml block solution).
24. Resuspend in 100 μl block solution.

2.2.3. Cell lysis and sonication

The cells are lysed to remove the bulk of cytosolic proteins, leaving only the contents of the nucleus for ChIP. This lysis step can improve ChIP results in cases where the protein of interest is not only bound to chromatin but also abundant in the cytosol. The successful isolation of nuclei can be confirmed after step 2 using standard Trypan blue staining. All lysis buffers should be supplemented with protease inhibitors (Complete, EDTA-free, Roche, #11873580001). Settings for the sonication of chromatin must be pre-determined based on equipment and material. The equipment and settings described here work well with most cell lines and primary tissues. After sonication, the opaque lysate should become clear as a first indicator of a successful sonication. If the lysate does not clear after additional cycles of sonication, the material may be excessively crosslinked and the crosslinking time in step 2 (for cells) or step 10 (for tissues) should be reduced. Ideally, most chromatin fragments resulting from sonication occur between 200 and 400 bp. This size range can be confirmed by running the whole-cell extract (WCE) on an agarose gel or an Agilent Bioanalyzer after reversing formaldehyde crosslinking and the DNA purification subsequent to step 54. (Fig. 2A).

25. Resuspend each pellet of crosslinked tissue in 10 ml of LB1 (50 mM Hepes–KOH, pH 7.5; 140 mM NaCl; 1 mM EDTA; 10% Glycerol; 0.5% NP-40 or Igpal CA-630; 0.25% Triton X-100). Rock at 4°C for 10 min. Spin at $2000 \times \text{rcf}$ for 4 min at 4°C in a tabletop centrifuge.
26. Resuspend each pellet in 10 ml of LB2 (10 mM Tris–HCl, pH8.0; 200 mM NaCl; 1 mM EDTA; 0.5 mM EGTA). Rock gently at 4°C for 5 min. Pellet nuclei in tabletop centrifuge by spinning at $2000 \times \text{rcf}$ for 5 min at 4°C .
27. Resuspend each pellet in each tube in 3 ml LB3 (10 mM Tris–HCl, pH 8; 100 mM NaCl; 1 mM EDTA; 0.5 mM EGTA; 0.1% Na–Deoxycholate; 0.5% *N*-lauroylsarcosine).
28. Transfer cells to a homemade “sonication tube” (cut a polypropylene, 15 ml conical tube into two pieces at the 7 ml mark).
29. Sonicate suspension with a microtip attached to a Misonix Sonicator 3000 Homogenizer sonicator. Samples should be kept in an ice-water bath during sonication. Sonicate 8–12 cycles of 30 s ON and 60 s OFF. Power-output should be between 27 and 33 Watt.

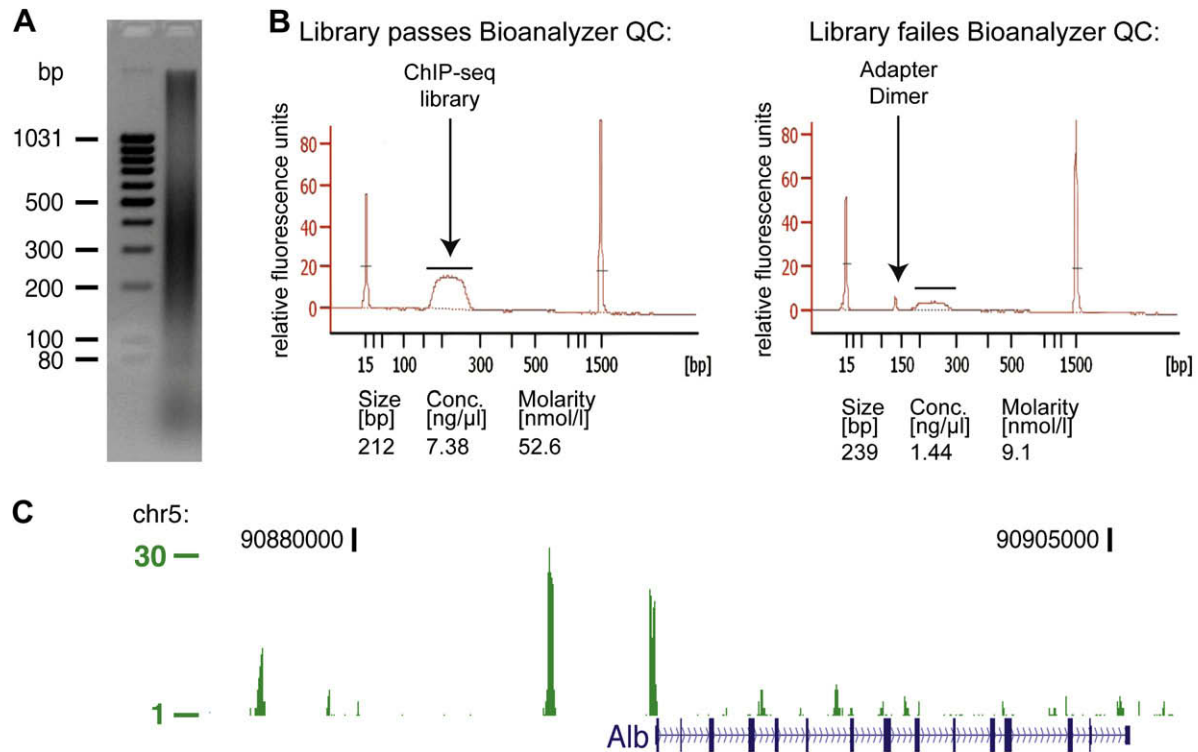


Fig. 2. (A) Example whole-cell extract (WCE) sonication result. (B) Agilent Bioanalyzer 2100 traces for two ChIP-seq libraries. The left panel shows a successful library preparation. The right panel shows a library with significant amounts of adapter dimers. The quantification of the libraries is shown underneath each panel. (C) C/EBP α ChIP-seq genome track (absolute fragment count) at the albumin locus in mouse hepatocytes showing several strong and weaker binding events.

30. Add 300 μ l of 10% Triton X-100 to sonicated lysate. Split into 2 ml centrifuge tubes. Spin at 20,000 \times rcf for 10 min at 4 $^{\circ}$ C to pellet debris.
31. Combine supernatants from the 2 ml centrifuge tubes in a new 15 ml conical tube. The amount of LB3 and Triton X-100 is adjusted to the number of ChIPs to be performed. For example to prepare 3 ChIPs from one 3 ml sonication you would top up the centrifuged sonication to 9 ml with LB3 and 600 μ l of 10% Triton X-100 (1% final concentration). Mix well and split into three 15 ml conical tubes, so that each contains 3 ml of cell lysate with 300 μ l Triton per ChIP.
32. Save 50 μ l of cell lysate from each sonication as whole-cell extract (WCE) DNA. Store at -20° C.

2.2.4. Chromatin immunoprecipitation

33. Add 100 μ l antibody/magnetic bead mix from step 24 to cell lysates.
34. Gently mix overnight on rotator or rocker at 4 $^{\circ}$ C.

2.2.5. Wash, elution, and crosslink reversal

Steps 35 through 40 should be done in a 4 $^{\circ}$ C cold room.

35. Pre-chill one 1.5 ml microfuge tube for each IP.
36. Transfer half the volume of an IP to a pre-chilled tube.
37. Let tubes sit in magnetic stand to collect the beads. Remove supernatant and add remaining IP. Let tubes sit again in magnetic stand to collect the beads and remove supernatant.
38. Add 1 ml RIPA Buffer (50 mM Hepes-KOH, pH 7.5; 500 mM LiCl; 1 mM EDTA; 1% NP-40 or Igelal CA-630; 0.7% Na-Deoxycholate) to each tube. Remove tubes from magnetic stand and shake or agitate tube gently to resuspend beads. Replace tubes in magnetic stand to collect beads. Remove supernatant. Repeat this wash 4–6 more times.

39. Wash once with 1 ml TBS (20 mM Tris-HCl, pH 7.6; 150 mM NaCl).
40. Spin at 960 \times rcf for 3 min at 4 $^{\circ}$ C and remove any residual TBS buffer using the magnetic stand.
41. Add 200 μ l of elution buffer (50 mM Tris-HCl, pH 8; 10 mM EDTA; 1% SDS).
42. Elute and perform reverse crosslinking at 65 $^{\circ}$ C for 6–18 h. Resuspend beads in the first 15 min with brief vortexing every 5 min.
43. Thaw 50 μ l of the WCE from step 32, add 150 μ l of elution buffer and mix. Reverse the formaldehyde crosslinking as in step 42 simultaneously with the ChIP samples.

2.2.6. Digestion of cellular protein and RNA

The proteins and RNA in the samples are enzymatically digested and the DNA is further purified by phenol-chloroform extraction and ethanol precipitation. GlycoBlue (Ambion, AM9516) can be used instead of glycogen as carrier for the ethanol precipitation, which substantially improves visualization of the DNA pellet.

44. Remove 200 μ l of supernatant and transfer to new tube.
45. Add 200 μ l of TE to each tube of IP and WCE DNA to dilute SDS in elution buffer.
46. Add 8 μ l of 1 mg/ml RNaseA (Ambion Cat # 2271).
47. Mix and incubate at 37 $^{\circ}$ C for 30 min.
48. Add 4 μ l of 20 mg/ml proteinase K (Invitrogen, 25530-049).
49. Mix and incubate at 55 $^{\circ}$ C for 1–2 h.
50. Add 400 μ l phenol-chloroform-isoamyl alcohol (P:C:IA) and separate phases with 2 ml Phase Lock Gel Light tubes FPR5101 Flowgen Bioscience and follow the instructions provided.

51. Transfer aqueous layer to new centrifuge tube containing 16 μ l of 5 M NaCl (200 mM final concentration) and 1 μ l of 20 μ g/ μ l GlycoBlue (Ambion, AM9516).
52. Add 800 μ l 100% EtOH. Incubate for 30 min at -80°C .
53. Spin at $20,000 \times \text{rcf}$ for 10 min at 4°C to pellet DNA. Wash pellets with 500 μ l of 80% EtOH and spin at $20,000 \times \text{rcf}$ for 5 min.
54. Dry pellets 10–20 min in a speedvac at 45°C and resuspend each in 30 μ l of 10 mM Tris–HCl, pH 8.0.
55. Measure DNA concentration of WCE with NanoDrop 1000 (Thermo Fisher Scientific). Note that ChIP samples are too low in DNA concentration to give reliable results using a NanoDrop.

2.2.7. Perform end-repair of the DNA

The final steps of this protocol convert the ChIP-enriched DNA into a library suitable for high-throughput sequencing using an Illumina Genome Analyzer. Historically, the Illumina Genomic Sample Preparation Kit has been used for ChIP reactions, since no dedicated ChIP-seq kit was available. Indeed, the recently released ChIP-seq kit is very similar to the Genomic Sample Preparation Kit. Here, we have replaced all the enzymes from the Illumina Genomic Sample Prep Kit using standard, commercially available products. Only the Adapter Oligonucleotide Mix and the PCR primers 1.1 and 2.1 should be directly ordered from Illumina, as they contain proprietary modifications that, based on our experience, greatly improve library synthesis.

Using commercial reagents as opposed to pre-assembled kits greatly reduces the price per library generation, and allows the preparation of a master mix for the following reactions. For new users, or if only a small number of samples are to be processed at a time, it may be simpler to use the Illumina Genomic Sample Preparation Kit. ChIP-seq libraries can also be prepared using paired-end adapters and PCR primers, as they are compatible with both single- and paired-end flowcells. However, we have not optimized the protocol for the paired-end adapters. Based on our prior experiences, we predict that concentration of the paired-end adapter in the ligation reaction will need to be optimized carefully.

56. Pipette the following mix into PCR tubes and keep on ice, or make a master mix on ice containing water, buffer and enzymes, and add to the samples. Incubate 30 min at 20°C in a thermal cycler.

ChIP sample, or 5–50 ng of WCE	30.0 μ l
Water	45.0 μ l
T4 DNA ligase buffer(NEB, B0202S)	10.0 μ l
dNTP mix, each 10 mM (NEB, N0447L)	4.0 μ l
T4 DNA polymerase (NEB, M0203L)	5.0 μ l
Klenow DNA polymerase (NEB, M0210L)	1.0 μ l
T4 PNK (NEB, M0201L)	5.0 μ l
Total	100.0 μ l

57. Clean-up samples using the DNA Clean&Concentrator-5 kit, (Zymo Research, USA), following the manufacturer's protocol.
58. Elute with 33 μ l EB preheated to 50°C . Chill on ice.

2.2.8. Add "A" bases to the DNA

Pipette the following mix in 1.5 ml tubes and keep on ice, or make a master mix on ice containing buffer and enzyme and add to the samples.

DNA sample	32.0 μ l
Klenow buffer (NEB, B7002S)	5.0 μ l
dATP (1 mM)	10.0 μ l
Klenow 3'–5' exo minus (NEB, M0212L)	3.0 μ l
Total	50.0 μ l

59. Incubate 30 min at 37°C in a water bath.
60. Clean-up samples using the DNA Clean&Concentrator-5 kit, following the manufacturer's protocol.
61. Elute with 9 μ l EB preheated to 50°C . Chill on ice.

2.2.9. Ligate sequencing adapters to DNA fragments

One of the most persistent problems we have observed is the formation of adapter dimers generated during adapter-target DNA ligations. Dimers form clusters on the flowcell of the Illumina Genome Analyzer and thus compete with the desired sample for sequencing. This can reduce the sequencing reads from the actual ChIP experiment. A number of steps can be taken to significantly reduce adapter dimers: (i) the amount of Adapter Oligonucleotide mix can be titrated by diluting the Adapter Oligonucleotide mix 40-fold. This gives robust results with as little as 5 ng of DNA; (ii) pooling of multiple ChIPs can be used to increase the relative amount of sample DNA versus Adapter Oligonucleotides; (iii) ultrapure ligases can be used, such as those from Enzymatics [25]; (iiii) after PCR amplification the library can be purified by solid-phase reversible immobilization technology as described in [25]; and (iv) Illumina recommends a gel purification step following the ligation reaction which will likely minimize these adaptor dimers but may result in loss of sample complexity in the case of ChIP-seq.

Pipette the following mix in 1.5 ml tubes on ice. Alternatively, add a master mix containing buffer and Adapter Oligo to the samples followed by the ligase.

DNA sample	8.0 μ l
Quick Ligation Reaction Buffer (NEB, M2200L)	12.5 μ l
Fourtyfold diluted Genomic Adapter Oligo mix (Illumina)	2.0 μ l
Quick T4 DNA Ligase (NEB, M2200L)	2.5 μ l
Total	25.0 μ l

62. Incubate 15 min at RT.
63. Clean-up samples using the DNA Clean&Concentrator-5 kit, following the manufacturer's protocol.
64. Elute with 24 μ l EB preheated to 50°C . Chill on ice.

2.2.10. Amplify adapter-modified DNA by PCR

ChIP-seq libraries at this stage have by nature only a small mass. To reduce the risk of complexity loss, we perform the PCR amplification *before* the size-selection step in agarose gel.

The library is amplified using a DNA polymerase that: (i) high fidelity and (ii) produces blunt ends. The recommended $2 \times$ master mix (NEB, F-531L) containing Phusion DNA polymerase should be distributed into convenient aliquots to avoid multiple freeze-and-thaw cycles. Alternatively, PCR yield can also be improved using Platinum Pfx polymerase (Invitrogen) see [25] for details.

Pipette the following mix directly into PCR tubes and keep on ice.

DNA sample	23.0 µl
Phusion Master Mix with HF Buffer (NEB, F-531L)	25.0 µl
Genomic PCR primer 1.1 (Illumina)	1.0 µl
Genomic PCR primer 2.1 (Illumina)	1.0 µl
Total	50.0 µl

65. Run program:

Step 1: 98 °C 30 s
 Step 2: 98 °C 10 s
 Step 3: 65 °C 30 s
 Step 4: 72 °C 30 s
 Step 5 GOTO step 2 17 times
 Step 6: 72 °C 5 min
 Step 7: 4 °C HOLD

66. Purify with QIAquick and elute with 32 µl preheated EB. This sample is called SolexaPreGel. If validation of ChIP-seq library is desired, follow the optional protocol for reamplification (steps 75–80). Alternatively purify using solid-phase reversible immobilization technology as described in [25].

2.2.11. Gel purification of SolexaPreGel for ChIPseq

The amplified library is purified on an agarose gel to select a specific size-range for cluster generation, as well as to remove potential adapter dimers. One sample per gel can be used to avoid crosscontamination of different libraries.

Using Xylene cyanol (Sigma, X4126) as loading dye has the advantage that it runs above the actual library, and does not interfere with the visualisation of the critical size range (150–700 bp) on a transilluminator.

67. Cast the appropriate number of 50 ml 2% agarose (Bio-Rad, 161-3106) TAE gels with 5 µl SybrSafe (S33102).
68. Add 3 µl of loading buffer (50% glycerol supplemented with 0.25% Xylene cyanol) to 8 µl of DNA ladder (NEB, N3233L).
69. Add 10 µl of loading buffer to each sample.
70. Load the entire ladder into the first well of the gel, leave one lane empty and load the sample into the next well. Load only one sample per gel to eliminate any possibility of crosscontamination.
71. Run gel at 120 V for 40 min.
72. Excise the 200–300 bp fragments on a Dark Reader (Claire Chemical Research) and purify the DNA with a Qiagen MinElute Gel Extraction Kit (Qiagen, 28606). When the DNA is extracted it might be advantageous *not* to heat the gel slice to 50 °C but to dissolve the gel slice at room temperature as discussed in [25]. Elute with 15 µl EB preheated to 50 °C. You can excise and store the larger fragments (300–800 bp) as a backup.
73. Run the library on a Bioanalyzer DNA 1000 assay (Agilent) to estimate the concentration and to check that no adapter dimers are present (Fig. 2B). If there are adapter dimers visible, the library could be rescued by solid-phase reversible immobilization technology as described in [25].
74. The sample is now ready for sequencing on an Illumina Genome Analyzer (Section 2.3).

2.2.12. Optional protocol for reamplification

We have described and validated an additional procedure that begins by diluting 1 µl of the amplified ChIP-seq library (SolexaPreGel) in 9 µl of EB buffer for subsequent analysis using real-time

PCR or ChIP-chip [14]. This approach allows direct testing of libraries or to confirm sequencing results with readily available technologies, such as DNA microarrays or real-time PCR. This portion of material should be set aside routinely. For this method, it is necessary to process a WCE sample at the same time as a reference for real-time PCR and/or ChIP-chip.

75. Use 2 µl of the diluted SolexaPreGel sample (1 µl SolexaPreGel in 9 µl EB) per PCR. One should also amplify the WCE sample that will be used as an input control for subsequent analyses.

76. Make PCR mix:

Stock	1 × Mix
10× Thermopol buffer (NEB)	5.0 µl
dNTP mix (25 mM each)	0.5 µl
Primer 1.2 for reamp ¹ (10 µM)	2.5 µl
Primer 2.2 for reamp ² (10 µM)	2.5 µl
AmpliTaq	1.0 µl
ddH ₂ O	36.5 µl
Total	48.0 µl

¹ Order the following primer: Primer 1.2 for reamplification: 5'AATGATACGGC GACCACCGAGATCTACACTCTTCCCTACACGACGCTCTCCGATCT.

² Order the following primer: Primer 2.2 for reamplification: 5'CAAGCAGAAGA CCGCATACGAGCTCTCCGATCT Oligonucleotide sequences reference: http://intro-n.cam.uchc.edu/groups/tgcore/wiki/013c0/Solexa_Library_Primer_Sequences.html.

77. Add 48 µl of PCR Mix to each sample and run program:

Step 1: 95 °C 2 min
 Step 2: 95 °C 30 s
 Step 3: 65 °C 30 s
 Step 4: 72 °C 1 min
 Step 5: GOTO step 2 24 times
 Step 6: 72 °C 5 min
 Step 7: 4 °C HOLD

77. After PCR is completed, clean-up samples with QIAquick minelute PCR Purification Kit. Elute with 25 µl EB.
78. Measure DNA concentration of all samples with NanoDrop.
79. Samples are ready for further processing towards microarray or real-time PCR.

2.2.13. Optional rescue of traditional ChIP-chip libraries for ChIP-seq

Ligation mediated PCR (LMPCR) [26] has been extensively used to amplify ChIP enriched DNA fragments [3,5,7,27]. Like the procedure presented here for building ChIP-seq libraries, LMPCR involves ligating annealed linkers to the DNA of interest, followed by a PCR amplification and (often) microarray analysis (ChIP-chip). While it is generally preferable to repeat ChIP-seq with new experiments, there are cases where the original material used for ChIP-chip was valuable enough to warrant recovery of the library; for instance, our laboratory uses primary human islets and hepatocytes samples that are difficult to obtain. While the original ChIP-chip libraries could be processed into a ChIP-seq library by addition of new linkers, the short nature of the reads currently obtained by high-throughput sequencing technology necessitates the removal of the original ChIP-chip linkers prior to library generation. Otherwise, the first 25 bases would consist of the original ChIP-chip linker sequence. Here, we present an optional procedure to remove a ChIP-chip linker from a previously made LMPCR library, so it can be

re-ligated to Illumina linkers, and used for ChIP-seq. While it is possible that additional amplifications during the Solexa library preparation could introduce bias into the results, we have obtained ChIP-seq results from ChIP-chip libraries that are fully consistent with the original ChIP-chip experiment. In order to control for any such bias, we recommend performing the same procedure on the LMPCR amplified input material that was used in the original ChIP-chip experiment.

81. Clean-up ChIP-chip library using the DNA Clean&Concentrator-5 kit, following the manufacturer's protocol. Elute in 42 μ l EB preheated to 50 °C.
82. Add 5 μ l of 10 \times PNK buffer (NEB, B0201S) to cleaned ChIP-chip library and heat to 70 °C for 10 min and chill on ice immediately (this may increase the efficiency of the subsequent phosphorylation step but can be omitted). Add 5 μ l (50 U) PNK (NEB, M0201S) and 5 μ l 10 mM ATP and incubate at 37 °C for 1–2 h.
83. Clean-up samples using the DNA Clean&Concentrator-5 kit, following the manufacturer's protocol. Elute in 9 μ l pre-warmed 50 °C EB.
84. Exonuclease digestion of linkers on phosphorylated LMPCR library: Mix directly on ice in a PCR tube: 8.5 μ l sample; 1 μ l of 10 \times reaction buffer (NEB, B0262S) and 0.5 μ l Lambda exonuclease (NEB, M0262S). Digest for 10 min at 37 °C in a PCR block. Heat inactivate for 10 min at 75 °C.
85. Clean-up samples using the DNA Clean&Concentrator-5 kit and elute in 30 μ l EB preheated to 50 °C.
86. Continue with ChIP-seq library generation (step 56–74).

2.3. Sequencing process

The Illumina Genome Analyser sequencing processes and biochemistry have been well described [28]. The sequencing capacity of next generation high-throughput sequencing machines is increasing at an almost exponential rate; for instance, the Illumina Genome Analyzer was able to produce 1 Gb of sequence per flow cell in January 2008, yet by December 2009, predicted yields are estimated to be 100 Gb. The amount of DNA sequence, the length of DNA reads and quality of the data produced by the next-generation sequencing technologies are only likely to increase.

Many improvements to the standard Illumina protocol [25,29] have been reported, most of which focus on the upfront sample preparation rather than the particular sequencing biochemistry. There are also ongoing developments in data analysis, primarily in genome alignment tools and peak calling, but also in image processing (see below). The most significant hurdle for efficient operation of the Genome Analysers involves quantification of the DNA library before cluster generation. The original methodologies of standard UV spectroscopy followed by titration of libraries to achieve optimal cluster density often afforded variable densities, and were laborious. The use of the Agilent Bioanalyser allows much better quality control of libraries prior to sequencing. In our hands, implementing the use of the Bioanalyser has increased the quality of library submissions to our sequencing service and improved the output of high quality sequencing data. The new high-sensitivity DNA1000 kit from Agilent improves detection of samples 20-fold. This allows quantification and QC of libraries from smaller amounts of starting material or fewer cycles of PCR amplification, both desirable to most users. The next generation technologies all require generation of a “sample prep spike” where minute quantities of DNA are prepared into libraries for sequencing, massively amplified for quantitation, and then massively diluted for sequencing. It would be preferable to bypass this amplification entirely and directly quantitate adapter ligated nucleic acid mole-

cules, opening the way to improved analysis of limited clinical or biological resources.

Quantitative real-time PCR [25] allows very robust quantitation and uniform cluster densities from Illumina ready libraries. The protocols are slightly complicated by the need for multiple primer-probe combinations as the adapter molecules are different for single end and paired end, or DNA and RNA libraries. This is being addressed by Illumina and standard adapter sequences are scheduled for release in 2009/10.

2.4. Quality control and data analysis

2.4.1. Basic data pipeline

The raw data format of the Illumina sequencer is images files. After each completed sequencing run these images are computationally processed to obtain nucleotide-base calls. Besides the standard analysis pipeline provided by Illumina, alternative base-calling algorithms exist, including Alta-Cyclic [30] and Rollexa [31], which are reported to reduce error rates and thus produce a higher number of alignable reads. Alternative programs tend to be more CPU intensive than the standard Illumina pipeline, a cost that somewhat counterbalances the sequence gains. However, improved base-calling will allow for longer and more reliable sequence reads and should in principle help map reads that cross into repetitive regions by anchoring them in the surrounding non-repetitive sequence. This procedure could also improve the reliability of the identification of single nucleotide polymorphisms (SNPs), and thus allele-specific protein–DNA contacts.

Subsequent to base-calling, the sequencing reads have to be aligned to a reference genome. Several applications are available to align the sequencing reads to a reference. Among many others there are ELAND [32], MAQ (maq.sourceforge.net) and Bowtie (bowtie-bio.sourceforge.net). The main differences among these algorithms are the use of quality values and the treatment of reads that map to multiple locations. MAQ uses the quality values provided by the base caller (which indicate the probability that the base is called correctly) to resolve mismatches in alignments. With MAQ, a mismatch at a low quality base is penalized less than a mismatch at a high quality base, since it is more likely that the difference is a sequencing error in that case. Bowtie and ELAND do not use quality values. If a read aligns to multiple positions in the reference genome equally well, MAQ and Bowtie choose one of those positions uniformly at random. In this case ELAND assigns these reads to an arbitrary, but not necessarily random, locations. Note that since MAQ uses quality values in scoring alignments and Bowtie does not, it is more likely that Bowtie will assign the same score to two alignments than will MAQ.

It should be noted that for the identification of binding events in some repetitive areas of the genome, the precise treatment of sequencing reads that map multiple times to the reference genome can be critical. However, these cases seem to represent a minority compared to the bulk of binding events.

2.4.2. Examination of aligned data as first quality control

In order to inspect if a ChIP-seq experiment was successful, it is convenient to view the alignment results as continuous-valued data in track formats, such as wiggle (WIG), GFF (General Feature Format), or bedGraph using for example the UCSC or Ensembl genome browser. Fig. 2C shows a wiggle track for a ChIP-seq experiment against the liver master regulator C/EBP α at the albumin locus performed in primary mouse liver. The height of the track represents the number of overlapping sequencing reads at bp resolution. This visualization allows a quick evaluation of the enrichments present in the data.

2.4.3. Automated identification of binding events

Following confirmation of successful genomic enrichment in a ChIP experiment, the next task is to identify the regions across the whole genome that are enriched in sequencing reads and thus harbor the DNA–protein interaction *in vivo*. Several algorithms have been developed to analyze ChIP-seq data and identify the locations of transcription factor binding sites and histone marks along the genome.

2.4.3.1. ChipSeq peak finder. ChipSeq Peak Finder [11] clusters the reads and uses the ratio of the counts in the immunoprecipitated and the control sample to call peaks. An updated version of the method, eRange [33], also allows the use of reads which map to multiple locations in the genome which results in a significant increase in the amount of data utilized.

2.4.3.2. XSET. The extended set method XSET [16] uses the full estimated length of the DNA fragments to call the regions with highest numbers of overlapping fragments.

2.4.3.3. Mikkelsen methodology. The method in Mikkelsen et al. [34] takes into account the ‘mappability’ of the underlying sequence, a measure of how many reads could be uniquely mapped at each location, and computes *p*-values to find significant differences between the observed and expected number of fragments.

2.4.3.4. PeakSeq. PeakSeq [35] allows for this mappability effect, which starts with a normalization step comparing the control with the background component of the ChIP sample and then detects significantly high concentrations of reads using the Binomial distribution.

2.4.3.5. MACS. Model-based Analysis for ChIP-seq (MACS) [36] shifts the tags on the forward and reverse strand together and uses the Poisson distribution to detect enrichment. In addition, the method ignores multiple identical reads to avoid biases during amplification and sequencing library preparation.

2.4.3.6. QuEST. Quantitative enrichment of sequence tags (QuEST) [37] shifts the peaks from opposite strands together and produces a kernel density estimation-derived score to call the enriched regions.

2.4.3.7. FindPeaks. FindPeaks [38] calls peaks according to some minimum height criteria without including a control sample in the analysis.

2.4.3.8. SISSR. Site Identification from Short Sequence Reads (SISSR) [39] estimates high read counts using Poisson probabilities and calls regions where the peaks shift from the forward to the reverse strand.

2.4.3.9. Other methods. In Kharchenko et al. [19] three similar peak calling methods are proposed, scoring read counts upstream and downstream of the each region to match tag patterns in the forward and reverse strands. In addition, Nix et al. [40] have simulated spike-in data, combined them with input reads from real experiments and used different metrics to score the peaks controlling for false discoveries. Another method that has been developed is BayesPeak which uses hidden Markov models and Bayesian techniques to identify the enriched regions based on posterior probabilities [41].

As with any new technology, it will take some time until the analysis of ChIP-seq experiments is a more standardized process. The growing number of tailored web-based tools and the advances made in sequencing throughput and quality will facilitate and im-

prove routine analysis in the future, and make this technology available to a broader group of researchers.

3. Concluding remarks

Using ChIP-seq, it is possible to ask, at a genome wide level, where and when proteins interact with DNA. As more high-throughput sequencers become available, the amount of information obtained through ChIP-seq is limited only by the available antibodies, sufficient starting material, and an accurate reference genome sequence on which to align results. The maps of transcription factor binding and modified histones generated by ChIP-seq are important resources for further functional investigation of the processes and mechanisms involved in gene regulation.

Acknowledgement

We thank J.S. Carroll for discussions, and N. Matthews for technical assistance and advice.

References

- [1] L. Elnitski, V.X. Jin, P.J. Farnham, S.J. Jones, Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques, *Genome Res.* 16 (2006) 1455–1464.
- [2] A. Rabinovich, V.X. Jin, R. Rabinovich, X. Xu, P.J. Farnham, E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites, *Genome Res.* 18 (2008) 1763–1777.
- [3] J.S. Carroll, X.S. Liu, A.S. Brodsky, W. Li, C.A. Meyer, et al., Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1, *Cell* 122 (2005) 33–43.
- [4] T.H. Kim, Z.K. Abdullaev, A.D. Smith, K.A. Ching, D.I. Loukinov, et al., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome, *Cell* 128 (2007) 1231–1245.
- [5] S. Cawley, S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, et al., Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs, *Cell* 116 (2004) 499–509.
- [6] M.D. Wilson, N.L. Barbosa-Morais, D. Schmidt, C.M. Conboy, L. Vanes, et al., Species-specific transcription in mice carrying human chromosome 21, *Science* 322 (2008) 434–438.
- [7] D.T. Odom, R.D. Dowell, E.S. Jacobsen, W. Gordon, T.W. Danford, et al., Tissue-specific transcriptional regulation has diverged significantly between human and mouse, *Nat. Genet.* 39 (2007) 730–732.
- [8] A.R. Borneman, T.A. Gianoulis, Z.D. Zhang, H. Yu, J. Rozowsky, et al., Divergence of transcription factor binding sites across related yeast species, *Science* 317 (2007) 815–819.
- [9] V. Orlando, Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation, *Trends Biochem. Sci.* 25 (2000) 99–104.
- [10] M.J. Solomon, P.L. Larsen, A. Varshavsky, Mapping protein–DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene, *Cell* 53 (1988) 937–947.
- [11] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions, *Science* 316 (2007) 1497–1502.
- [12] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, et al., High-resolution profiling of histone methylations in the human genome, *Cell* 129 (2007) 823–837.
- [13] B. Wold, R.M. Myers, Sequence census methods for functional genomics, *Nat. Methods* 5 (2008) 19–21.
- [14] D. Schmidt, R. Stark, M.D. Wilson, G.D. Brown, D.T. Odom, Genome-scale validation of deep-sequencing libraries, *PLoS ONE* 3 (2008) e3713.
- [15] G. Bourque, B. Leong, V.B. Vega, X. Chen, Y.L. Lee, et al., Evolution of the mammalian transcription factor binding repertoire via transposable elements, *Genome Res.* 18 (2008) 1752–1762.
- [16] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, et al., Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nat. Methods* 4 (2007) 651–657.
- [17] O. Morozova, M.A. Marra, Applications of next-generation sequencing technologies in functional genomics, *Genomics* 92 (2008) 255–264.
- [18] D.R. Smith, A.R. Quinlan, H.E. Peckham, K. Makowsky, W. Tao, et al., Rapid whole-genome mutational profiling using next-generation sequencing technologies, *Genome Res.* 18 (2008) 1638–1642.
- [19] P.V. Kharchenko, M.Y. Tolstorukov, P.J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins, *Nat. Biotechnol.*, 2008.
- [20] T.I. Lee, S.E. Johnstone, R.A. Young, Chromatin immunoprecipitation and microarray-based analysis of protein location, *Nat. Protoc.* 1 (2006) 729–748.

- [21] L.P. O'Neill, B.M. Turner, Immunoprecipitation of native chromatin, in: *NChIP*, *Methods* 31 (2003) 76–82.
- [22] L.G. Acevedo, A.L. Iniguez, H.L. Holster, X. Zhang, R. Green, et al., Genome-scale ChIP-chip analysis using 10,000 human cells, *Biotechniques* 43 (2007) 791–797.
- [23] L.P. O'Neill, M.D. VerMilyea, B.M. Turner, Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations, *Nat. Genet.* 38 (2006) 835–841.
- [24] T.H. Kim, B. Ren, Genome-wide analysis of protein–DNA interactions, *Annu. Rev. Genomics Hum. Genet.* 7 (2006) 81–102.
- [25] M.A. Quail, I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, et al., A large genome center's improvements to the Illumina sequencing system, *Nat. Methods* 5 (2008) 1005–1010.
- [26] P.R. Mueller, B. Wold, In vivo footprinting of a muscle specific enhancer by ligation mediated PCR, *Science* 246 (1989) 780–786.
- [27] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, et al., Genome-wide location and function of DNA binding proteins, *Science* 290 (2000) 2306–2309.
- [28] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, et al., Accurate whole human genome sequencing using reversible terminator chemistry, *Nature* 456 (2008) 53–59.
- [29] M. Meyer, A.W. Briggs, T. Maricic, B. Hober, B. Hoffner, et al., From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing, *Nucleic Acids Res.* 36 (2008) e5.
- [30] Y. Erlich, P.P. Mitra, M. delaBastide, W.R. McCombie, G.J. Hannon, Alta-cyclic: a self-optimizing base caller for next-generation sequencing, *Nat. Methods* 5 (2008) 679–682.
- [31] J. Rougemont, A. Amzallag, C. Iseli, L. Farinelli, I. Xenarios, et al., Probabilistic base calling of Solexa sequencing data, *BMC Bioinform.* 9 (2008) 431.
- [32] A.J. Cox, Ultra high throughput alignment of short sequence tags, in preparation.
- [33] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat. Methods* 5 (2008) 621–628.
- [34] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature* 448 (2007) 553–560.
- [35] J. Rozowsky, G. Euskirchen, R.K. Auerbach, Z.D. Zhang, T. Gibson, et al., PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls, *Nat. Biotechnol.* 27 (2009) 66–75.
- [36] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoutte, D.S. Johnson, et al., Model-based analysis of ChIP-Seq (MACS), *Genome Biol.* 9 (2008) R137.
- [37] A. Valouev, D.S. Johnson, A. Sundquist, C. Medina, E. Anton, et al., Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data, *Nat. Methods* 5 (2008) 829–834.
- [38] A.P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, et al., FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology, *Bioinformatics* 24 (2008) 1729–1730.
- [39] R. Jothi, S. Cuddapah, A. Barski, K. Cui, K. Zhao, Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data, *Nucleic Acids Res.* 36 (2008) 5221–5231.
- [40] D.A. Nix, S.J. Courdy, K.M. Boucher, Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks, *BMC Bioinform.* 9 (2008) 523.
- [41] C. Spyrou, personal communication.

Research article

Open Access

The pitfalls of platform comparison: DNA copy number array technologies assessed

Christina Curtis^{*†1,2}, Andy G Lynch^{*†1,2}, Mark J Dunning², Inmaculada Spiteri², John C Marioni³, James Hadfield², Suet-Feung Chin^{1,2}, James D Brenton^{1,2}, Simon Tavaré^{1,2} and Carlos Caldas^{1,2}

Addresses: ¹Department of Oncology, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB20XZ, UK, ²Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB20RE, UK and ³Department of Human Genetics, University of Chicago, 920 E 58th Street, Chicago, IL 60637, USA

E-mail: Christina Curtis^{*} - christina.curtis@cancer.org.uk; Andy G Lynch^{*} - andy.lynch@cancer.org.uk; Mark J Dunning - mark.dunning@cancer.org.uk; Inmaculada Spiteri - inma.spiteri@cancer.org.uk; John C Marioni - marioni@uchicago.edu; James Hadfield - james.hadfield@cancer.org.uk; Suet-Feung Chin - suet-feung.chin@cancer.org.uk; James D Brenton - james.brenton@cancer.org.uk; Simon Tavaré - simon.tavare@cancer.org.uk; Carlos Caldas - carlos.caldas@cancer.org.uk

^{*}Corresponding author [†]Equal contributors

Published: 8 December 2009

Received: 18 June 2009

BMC Genomics 2009, **10**:588 doi: 10.1186/1471-2164-10-588

Accepted: 8 December 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/588>

© 2009 Curtis et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The accurate and high resolution mapping of DNA copy number aberrations has become an important tool by which to gain insight into the mechanisms of tumourigenesis. There are various commercially available platforms for such studies, but there remains no general consensus as to the optimal platform. There have been several previous platform comparison studies, but they have either described older technologies, used less-complex samples, or have not addressed the issue of the inherent biases in such comparisons. Here we describe a systematic comparison of data from four leading microarray technologies (the Affymetrix Genome-wide SNP 5.0 array, Agilent High-Density CGH Human 244A array, Illumina HumanCNV370-Duo DNA Analysis BeadChip, and the Nimblegen 385 K oligonucleotide array). We compare samples derived from primary breast tumours and their corresponding matched normals, well-established cancer cell lines, and HapMap individuals. By careful consideration and avoidance of potential sources of bias, we aim to provide a fair assessment of platform performance.

Results: By performing a theoretical assessment of the reproducibility, noise, and sensitivity of each platform, notable differences were revealed. Nimblegen exhibited between-replicate array variances an order of magnitude greater than the other three platforms, with Agilent slightly outperforming the others, and a comparison of self-self hybridizations revealed similar patterns. An assessment of the single probe power revealed that Agilent exhibits the highest sensitivity. Additionally, we performed an in-depth visual assessment of the ability of each platform to detect aberrations of varying sizes. As expected, all platforms were able to identify large aberrations in a robust manner. However, some focal amplifications and deletions were only detected in a subset of the platforms.

Conclusion: Although there are substantial differences in the design, density, and number of replicate probes, the comparison indicates a generally high level of concordance between

platforms, despite differences in the reproducibility, noise, and sensitivity. In general, Agilent tended to be the best aCGH platform and Affymetrix, the superior SNP-CGH platform, but for specific decisions the results described herein provide a guide for platform selection and study design, and the dataset a resource for more tailored comparisons.

Background

The accurate and high-resolution mapping of DNA copy number aberrations (CNA) has become an important tool for biological and medical research. From understanding the extent of natural genetic variation [1], to associations with diseases such as HIV [2], to elucidating the mechanisms of tumorigenesis [3], such research is dependent on the quality of the data generated.

Numerous reports on the use and comparison of copy number profiling platforms have appeared [4-10] and more recently an approach to perform meta-analyses across such platforms has been described [11]. Early studies [12] suggested a high level of concordance between BAC-based aCGH and SNP-based platforms (Affymetrix 10 K array) in detecting CNA, but did not formally compare them. Greshock *et al.* [5] performed the first systematic comparison of multiple platforms on melanoma cell lines and found that a high level of sensitivity and specificity was observed for the Agilent 185 K arrays and that the increased probe density of Affymetrix arrays (100 K and 500 K) results in increased confidence in detection for these platforms. These results were echoed by Gunnarsson *et al.* [8] who also examined the performance of several older copy number profiling platforms (a 32 K BAC array, the Affymetrix 250 K SNP array, the Agilent 185 K oligonucleotide array, and the Illumina 317 K SNP) array in 10 chronic lymphocyte leukaemia (CLL) samples. They concluded that all platforms performed reasonably well at detecting large alterations, but that BAC probes were too large to detect small alterations. While Agilent offered the highest sensitivity, the increased density of SNP-CGH platforms (Affymetrix and Illumina) compensated for their increased technical variability, with Affymetrix detecting a higher degree of CNA compared to Illumina. A further aCGH study did not compare platforms, but did investigate the influence of cellularity on copy number detection [13] and concluded that modern high-resolution arrays could cope with high levels of contamination.

To attempt a fair and formal comparison of copy-number profiling platforms in a general setting is an almost futile exercise. Quantification of performance is difficult even with idealized data, and while measurements have been proposed such as the theoretical power to discover a single copy loss or gain [7], or the 'functional resolution' of the platform [6], these tend either to measure a very specific

aspect of the platform, or appear flawed under close examination. Such idealized data are, in any case, difficult to obtain, as one has to ask what is fair in terms of numbers entering the experimental design. Should one Illumina array be compared to one Nimblegen array or should the two-channel Nimblegen array be compared to two arrays from the single colour technology? Should the two-colour platform be penalized by an inefficient design to allow easier comparison, or the SNP-based platform credited for the additional information that it brings? If, as often is the case, the main experimental constraint is financial, then comparing \$1000 of one technology to \$1000 of another technology would seem sensible. However, the relative costs of platforms will vary from laboratory to laboratory and with time, and such an approach would foist the authors' view of microarray economics on the reader.

Additionally, the results from such an exercise are only as good as the analysis methods used and in that regard one has two options, both flawed. Naturally, the platforms will require different pre-processing strategies, but if different methods of analysis are also used for segmentation, then the performance of the technology will be confounded with the adequacy of the algorithm. This then punishes newer technologies for which analytical methodologies are not yet mature. The alternative, to use a common approach for the analysis of all platforms, is undesirable firstly because that approach is likely to have been developed for one of the technologies and may thus introduce bias, and secondly because the deliberate use of a sub-optimal analysis does not provide useful information to inform decisions in the real world. Nonetheless, informative qualitative comparisons can be made without performing segmentation that illuminate the relative strengths and weaknesses of each platform. We acknowledge that some users will be primarily interested in a comparison based on using existing analytical tools, rather than concerning themselves with the potential of each platform, but that is not the purpose of this study.

This study differs from previous comparative assessments of copy number profiling platforms in that we have attempted to characterize the strengths and weaknesses of various platforms in as unbiased a fashion as possible by avoiding measures that cannot be fairly computed, highlighting areas of potential bias, and

emphasizing a graphical assessment of performance that provides insight about the underlying technology as well as the specific platform. Inevitably, despite considerable effort, these comparisons will be shaped by our own prejudices concerning copy number analysis, but we have made the raw data available for others to draw their own conclusions.

Due to the speed of platform development, it is typical for a platform to be superseded by one with a greater number of features before comparisons involving it are published. The generation of platforms described here have not yet been the subject of an in-depth comparison, but have indeed already been superseded since this study was performed. Nonetheless, the underlying technologies are similar and a comparison is still informative. Implications for the new generations are discussed in the New Platforms section.

Herein we describe a comparison based on the analysis of two cell lines, six primary breast tumours, including matched normal samples, and two HapMap individuals. The SUM159 and MT3 cell lines and HapMap samples were selected based on the presence of known chromosomal aberrations, while the tumours are highly heterogeneous and hence present additional complexity for copy number analysis, not least with regard to their varying degrees of cellularity.

Here we present an analysis of probe coverage on each of the microarray platforms and a technical description of their reproducibility, sensitivity, and noise. We also provide an in-depth visual assessment of the ability of the different platforms to identify a range of sizes of copy number aberration. Lastly, we provide a publicly available dataset resulting from the processing of a range of samples (chosen to evaluate different abilities) on each platform. This information will allow interested parties to make decisions based on their own circumstances, preferences, and constraints.

Results

Theoretical and technical performance

Probe coverage and resolution

We present a summary of probe numbers in Table 1. Appreciation of the basic differences between the platforms is crucial for understanding the differences in performance. The Affymetrix platform has by far the most features, with the Illumina and Nimblegen arrays having a little under half of that number, and the Agilent array having markedly fewer still. More detailed summaries, including range of coverage and breakdown by chromosomal arm are presented in Additional File 1.

We choose not to present the theoretical functional resolution of these platforms as calculated by ResCalc [6]

Table 1: Basic summary of platform contents

Chromosome	Affymetrix	Agilent	Illumina	Nimblegen
1	64442	17259	27151	30220
2	69304	17382	28903	32900
3	58067	14802	24393	27255
4	55531	12863	22136	25940
5	52788	12486	22016	24223
6	51362	12438	26824	23138
7	43909	12201	20022	20549
8	45407	10309	20369	19870
9	34991	8461	17551	15160
10	42890	10297	18063	17820
11	41597	10114	16916	17901
12	40517	10169	16965	17991
13	30495	7375	13134	13541
14	25712	7512	11140	12130
15	23131	7314	10540	10735
16	22875	5610	10454	10206
17	19375	6220	9990	10025
18	24091	5586	11407	10682
19	12122	4081	7251	6828
20	19498	4715	8659	8403
21	11510	3077	5982	4733
22	10590	3181	6209	4442
X	27536	10179	12556	19151
Y	996	1191	1412	1963
Total	828736	214822	370043	385806

For the probes used in this study (i.e., only 60-mer Agilent probes, and well-annotated Affymetrix probes), the number of features broken down by chromosome.

for three reasons, each of which is, in itself, revealing with regard to the inter-platform differences. Firstly, the results presented in Coe *et al.* [6] obscure a large degree of inter- and intra- chromosome variability. As a proportion of their total, Illumina have more probes on chromosome 6 than do the other platforms, with the result that even though there are more probes in total on the Nimblegen platform, for this particular chromosome Illumina have 16% more probes than Nimblegen. On chromosome 19, Affymetrix put a noticeably higher

proportion of probes on the q arm than do Agilent, a situation that is reversed on chromosome 7.

The second problem of comparing by ResCalc is that the tool allows the platforms to define their own range of coverage from telomere to centromere. This makes it possible for a platform to improve its functional resolution by removing probes (essentially by dropping peripheral loosely spaced probes, while retaining the central tightly spaced ones), which is undesirable. To take an example, on arm 7p, in the core region covered by all of the platforms, Affymetrix average a probe every 3 to 4 Kb. However on the telomeric side of that core region, they have two probes covering 80 Kb. Undoubtedly the functional resolution as calculated by ResCalc would improve if such probes were removed (indeed, in this example, the removal of a single telomeric probe improves the reported functional resolution by 140 bases). Taking a more extreme example, the p arm of chromosome 9 has 13,643 probes on the Affymetrix platform and has a reported single probe functional resolution of 222,000 bases, but by removing 6 extreme telomeric probes and 166 extreme centromeric probes that are more sparsely positioned, we can improve the reported resolution to 8,900 bases. In general, the SNP-based platforms cover a wider region, with Nimblegen coming third and Agilent, in effect, often defining the core region of common coverage.

Finally, the hypothesis of uniform occurrence of CNA is doubtful and some of the platforms have been designed to provide non-uniform coverage by tiling more probes in known regions of variation (see Methods section for further details), or in areas where variation would be of particular interest. For example, Nimblegen have chosen, for the second generation of the product featured here, to switch from a uniform spacing along the genome to a 'designed' layout. This move would appear detrimental using tools such as ResCalc, but is clearly done for a purpose.

Reasons that one might adopt a non-uniform spacing include the desire to incorporate prior knowledge of genomic structure (e.g. to target CNVs, promoter regions, genes etc. and avoid repetitive elements), empirical evidence of probe performance from previous array designs, and lastly to achieve uniformity of probe performance. We show in the Methods section that there are a number of probe properties (most notably GC content) that affect the consistency of probe performance. These trends were visible in our data for all four platforms. There may, of course, be effects that are less visible, from these data, such as saturation levels and dynamic ranges. Naturally, increased probe coverage can address issues of variation, but technical biases will not be salved by increasing the number of probes.

Replicate probes

All of the platforms provide some replicate probes, by which we mean probes carrying the same sequence. For the SNP-CGH arrays, this is an integral aspect of the platforms and nearly all of the observations are actually averaged from replicate probes, 4 replicates for the Affymetrix SNP probes, and an average of 16 replicates for Illumina probes (although this ranges from 0 to over 40). With the Agilent and Nimblegen arrays, such probes are a rarity, and the majority of observations are based on only one probe. As such, for these two platforms, it makes sense to use the few probes with replicate information to characterize the performance of all observations. We can do this most informatively by calculating the variance of the replicate log-ratios between two samples.

Agilent provide, in addition to control probes, 916 60-mer probes for which there are three replicates. Nimblegen do not nominally provide any replication, but the coverage of the pseudoautosomal regions of the X and Y chromosomes results in 314 probes that are apparently replicated. However, we should note that these probes are treated as lying on different chromosomes, and thus if any within-chromosome normalization has taken place then their apparent reproducibility will be adversely affected. Neither Agilent nor Nimblegen show a strong association between the magnitude of log-ratio and variance of replicate observations (this is after all one of the reasons for analysing the log-ratio). To enable between-array comparisons, when we have resisted performing between-array normalizations, we summarize for the HapMap-HapMap comparisons the variance of replicate probes scaled by the mean difference in log-ratios observed in chromosomes X and 13, a difference that should be 1 for this comparison. Since this scaling does not share information between arrays, it is not a between-array normalization method.

For Agilent, the median variance of replicate probes is 0.042, 0.048, and 0.058 on three different arrays with third quartile values of 0.087, 0.111, and 0.120 respectively. In contrast, for Nimblegen, the median variance of replicate probes is 0.125, 0.142, and 0.144 with third quartile values of 0.309, 0.429, and 0.504, respectively. Thus Nimblegen exhibits 2-4 fold greater variability amongst replicate probes than Agilent. However, we note that the interpretation of the third quartile, in particular, should be tempered by our knowledge of the autocorrelation of probes along the genome.

Note that while the SNP-CGH platforms enable the quantification of allele-specific copy number [14-16], similar results cannot be obtained for the aCGH platforms. As such, we will focus strictly on the analysis of total copy number values. To quantify DNA abundance (or raw total copy number), the SNP-CGH platforms essentially sum the

fluorescence intensities from the two alleles investigated for a given SNP. This involves, for each allele, averaging over the replicate probes and then summing.

Because of these replicate probes, for Affymetrix and Illumina estimating the variance of individual probes is of limited value, since the values of individual probes will not be reported. Yet, for Illumina we cannot provide a good estimate of the variance after averaging over the replicate probes and then summing over alleles because the covariance of the two channels is not estimable from the data provided by *BeadStudio* [17], but can be presumed not to be zero due to the array design.

Replicate arrays

After scaling within arrays to obtain a difference of 1 for the chromosome X to chromosome 13 comparison, the variances of three replicate HapMap-HapMap comparisons were calculated. As can be seen in Table 2, Nimblegen exhibits between replicate array variances an order of magnitude greater than the rest.

Self-self comparisons

The ability of a copy number profiling platform to detect aberrations is largely determined by the noise observed in the measurements from that platform. This is a measure not only of the variance of the noise (although this is important), but also the kurtosis of the noise (i.e., if the

Table 2: Variance among three replicate HapMap-HapMap comparisons

Platform	1 st Quart	Median	Mean	3 rd Quart
Affymetrix	0.067	0.173	0.356	0.391
Agilent	0.046	0.122	0.304	0.284
Illumina	0.058	0.151	0.372	0.352
Nimblegen	1.21	3.03	5.65	6.47

After calculating the variance of each feature from three suitable scaled HapMap/HapMap comparisons, the mean, and quartile values of the variances are presented.

Table 3: Characteristics of a surrogate self-self hybridization

	Variance (scaled)	Autocorrelation	% z > 2	% z > 3
Affymetrix	0.33, 0.34, 0.29	0.040, 0.039, 0.036	4.5, 4.8, 4.7	1.3, 1.4, 1.3
Agilent	0.22, 0.24, 0.21	-0.001, 0.027, 0.019	4.4, 4.6, 4.9	0.5, 0.8, 0.7
Illumina	0.28, 0.36, 0.31	0.086, 0.066, 0.076	5.2, 5.3, 5.3	1.4, 1.4, 1.2
Nimblegen	0.81, 0.85, 0.60	0.009, 0.035, 0.026	4.3, 4.8, 4.5	0.5, 0.5, 0.5

For chromosome 13 of a HapMap/HapMap comparison (a surrogate self-self hybridization), presented are the variance, autocorrelation, and percentage of observations beyond two or three standard deviations.

noise is relatively heavy tailed, then more false calls will be made) and the independence of neighbouring probes. Not only are there known autocorrelation effects along the genome [18], possibly driven or exacerbated by autocorrelation in the quality of probe design caused by regions of high GC content or highly repetitive elements, but if probes are too close then they may compete to register the same DNA fragments. In such a case, the lack of independence of measurements from the probes would detract from the benefits of having improved probe density.

The ideal test for such a comparison would be a set of log-ratios generated from two replicate normal samples, as any departure from a log-ratio of 0 for these platforms must be noise and can be easily quantified. Since for two platforms, one of the pooled normal samples intended for this task was of lower quality, instead we again use chromosome 13 from a comparison of the two HapMap samples. Not only does this have no known changes, but adds the benefit that again we can scale our observations so that the difference in log-ratios between chromosomes 13 and X is a standard 1.

We summarize the noise by four measures in Table 3: the variance (after scaling as described), the autocorrelation of measurements at lag 1 along the chromosome, the percentage of observations beyond two standard deviations, and the percentage of observations beyond three standard deviations. The first measure will ideally be low and gives an indication of the noise-to-signal ratio, the second gives a measure of the independence of neighbouring probes, while the third and fourth give an idea of the false calling rates that might arise.

These results indicate that Nimblegen is noisy, exhibiting poor variance (2-4 fold greater than the other platforms). Additionally, Illumina has relatively poor autocorrelation for its probe density and has more outliers at a standard deviation of 2. Further, both SNP-CGH platforms have more outliers beyond a standard deviation of 3, which may be related to the autocorrelation. It is worth noting that Agilent has relatively few probes on chromosome 13 (see Table 1, Additional File 1), but

based on other performance measures, this is unlikely to influence significantly its superior performance.

Male-Female comparisons based on X and Y chromosomes

Since the two HapMap samples consist of a male (NA10851) and a female (NA15510), but for the autosomal chromosomes exhibit few copy number differences, we can use these samples to investigate the ability of a single probe on these platforms to distinguish between the diploid state and an altered copy number state due to regions of physical loss. We compare the log-ratios arising from chromosome 13 with those arising from chromosome X in order to test the ability to detect a 2:1 copy number alteration, and also with those arising from chromosome Y in order to test the ability to detect a 1:0 copy number alteration. The single-probe abilities of the four platforms are depicted in Figure 1.

For distinguishing between sites where both samples have two copies and sites where one sample has two copies while the other has one (13 versus X), Affymetrix and Agilent marginally outperform Illumina, while Nimblegen performs noticeably worse. In contrast, when distinguishing between sites where both samples have two copies and sites where one sample has no copies while the other has one (chromosome 13 versus Y), Agilent generally exhibits the highest sensitivity, although Illumina outperforms Agilent if very high specificity is sought. These are followed

in performance by Nimblegen, with Affymetrix performing considerably worse.

Notably, the Affymetrix Human Mapping 100 K, 500 K, and SNP5 platforms include chromosome X SNPs but no chromosome Y or mitochondrial SNPs. With the SNP5 platform, copy number non-polymorphic (CN probes) were introduced and for the Y chromosome there are 996 such probes with sufficient genomic information (1994 in total) all of which map outside the pseudoautosomal region. As such, for the SNP5 platform, the Y chromosome is not representative of other chromosomes in that it does not include any SNP probes and contains 0.1% of all probes on the platform. The lack of SNP probes is one possible explanation for the poor discrimination of a single copy loss on the Y chromosome. As noted in the Methods section, the CN probes are generally unreplicated and while few in number, the actual number of probes is on par with the other platforms.

Qualitative assessment of copy number aberration detection

The platforms investigated in this study differ substantially in their design, the number of probes, and their experimental utility. To obtain an overview of platform performance, the ability to detect several types of common chromosomal changes was assessed. In particular, the following alterations were considered based on

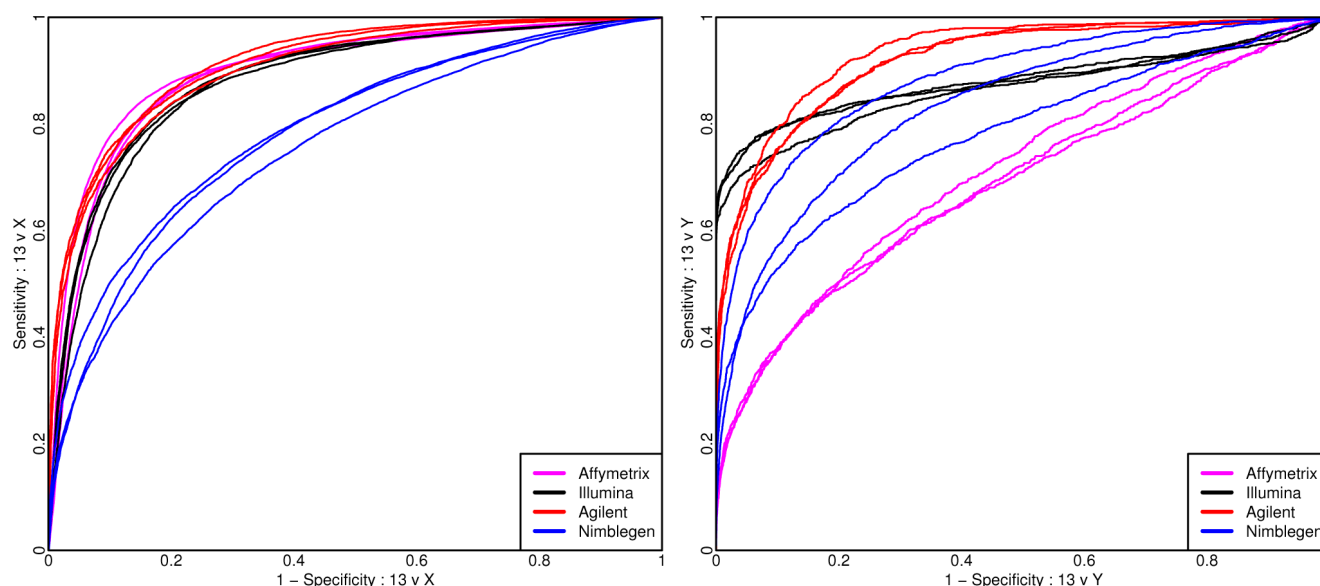


Figure 1

For a comparison of the HapMap samples ROC curves are presented to assess the performance of a single probe/probe-set for distinguishing the log-ratios associated with differing copy numbers from the log-ratios of chromosome 13 where copy-numbers should agree. Note the contrast from the left-hand panel, where the performances of Affymetrix and Agilent are indistinguishable, and the right hand panel, where the performance of Affymetrix has substantially declined.

raw copy number changes: whole chromosome gains or losses, chromosome arm gains or losses, high amplitude focal amplifications as well as subchromosomal gains and losses, small regions of gain or loss as exemplified by normal copy number variation.

i.) Whole chromosome gains or losses

This simple type of genomic aberration allows for examination of consistency at the level of probe log-ratios (or potentially segmented means) along the whole chromosome. Note that this is similar to the comparison of the HapMap samples in the male-female comparison. Here we use the MT3 cell-line, which is known to have single-copy gains of chromosomes 7 and 13, and a single copy loss of chromosome X. As would be expected, all four platforms can identify whole-chromosome events (Figure 2), but there are differences in the abilities to quantify the change and also in the discrimination of different copy number states that will be influential for the classification of smaller regions. Agilent performs best on both of these measures. Nimblegen includes probes targeting the pseudoautosomal region, which explains the apparent departure from zero for chromosome Y.

Also of note is the performance in terms of Y chromosome detection and the effect of normalization on the Illumina array. The performance of Illumina in detecting the absence of the Y chromosome in females is of concern. It is not unreasonable that what would ideally be an estimate of $\log_2(0/0)$ should be unstable (although due to non-specific binding the extremes of this instability will not be observed). If the observed values are indicative of any bias in the probe design, then the apparently strong performance of Illumina in the chromosome Y versus chromosome 13 comparison may have been misleading.

ii.) Chromosome arm gains or losses

We illustrate the ability of the platforms to detect a gain on a single arm of chromosome 5 in the SUM159 cell-line where, in addition to other variations, the 5p arm has an extra copy. Figure 3 illustrates the performance of the platforms for this chromosome. All of the platforms are able to detect the alteration, manifested as an upward deflection, but the clarity of signal is greatest for Agilent, followed by Affymetrix, Illumina and Nimblegen. This region is depicted in greater detail in Additional File 2. Of note is the duplication visible only in Illumina, at about 70 Mb into the chromosome. This is an area of known intra-chromosomal segmental duplication [2] and the other platforms place few probes in this region, as it is difficult to tile in these regions.

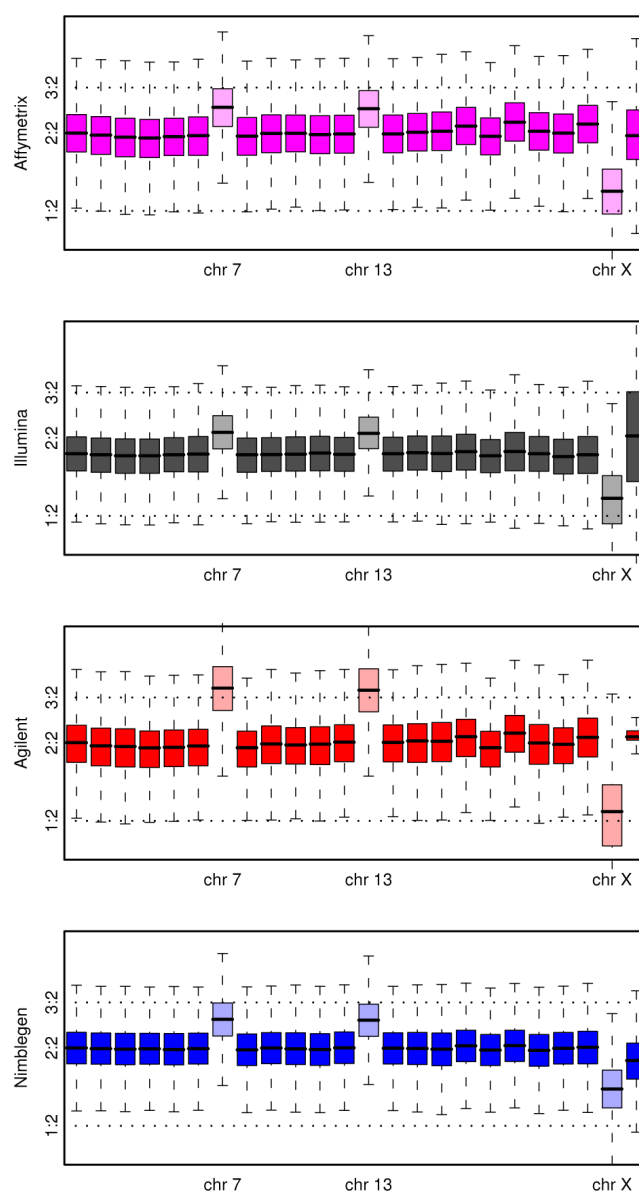


Figure 2
Showing, for a comparison of the MT3 cell-line to a pooled normal reference, a boxplot of the log-ratios from each platform broken down by chromosome.

Also indicated are theoretical markers for a single copy gain and a single copy loss. The three chromosomes with known aberrant copy number are indicated.

iii.) High amplitude focal amplifications and subchromosomal gains and losses

These smaller variants are relatively complex aberrations and test the abilities of the platforms to determine breakpoints accurately. These types of alterations would also allow for the easiest assessment of segmentation algorithms, if such a task were desired. Three examples occur on chromosome 5 of the SUM159 cell-line (Figure 3). The most

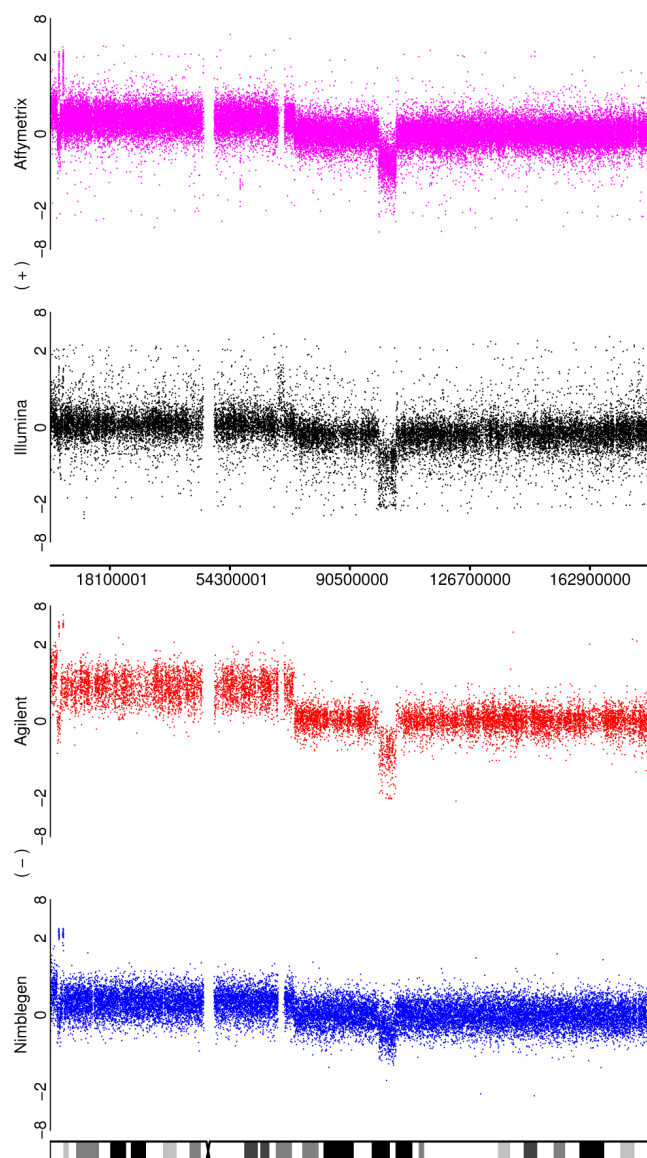


Figure 3
Illustrating the ability of the platforms to detect the duplication of a chromosomal arm. Depicted are the log-ratios for a comparison of the SUM159 cell line to a pooled normal reference for chromosome 5. In addition to a number of smaller aberrations, there is a duplication of the p arm of the chromosome for this sample.

obvious alteration (a deletion at approximately 100 Mb) is clearly observed in all four platforms, although again the difference is less obvious for Nimblegen. The second aberration, a complex change towards the telomere of arm 5p, is also seen by all four platforms, but the clarity of the pattern is variable. Once again, Agilent is generally clearest, but the two amplified regions are seen more clearly by Nimblegen than by the two SNP arrays, although they would still be detected by those platforms. The deletions

follow the usual order of being the most clear for Agilent > Affymetrix > Illumina > Nimblegen. The third much smaller change is most obvious for Affymetrix at about 55 Mb and is just barely detectable with Illumina, being so narrow as to fall between probes for Agilent and Nimblegen. A similar pattern is seen for the change at approximately 130 Mb on chromosome 8 for the SUM159 cell-line (Figure 4). Again,

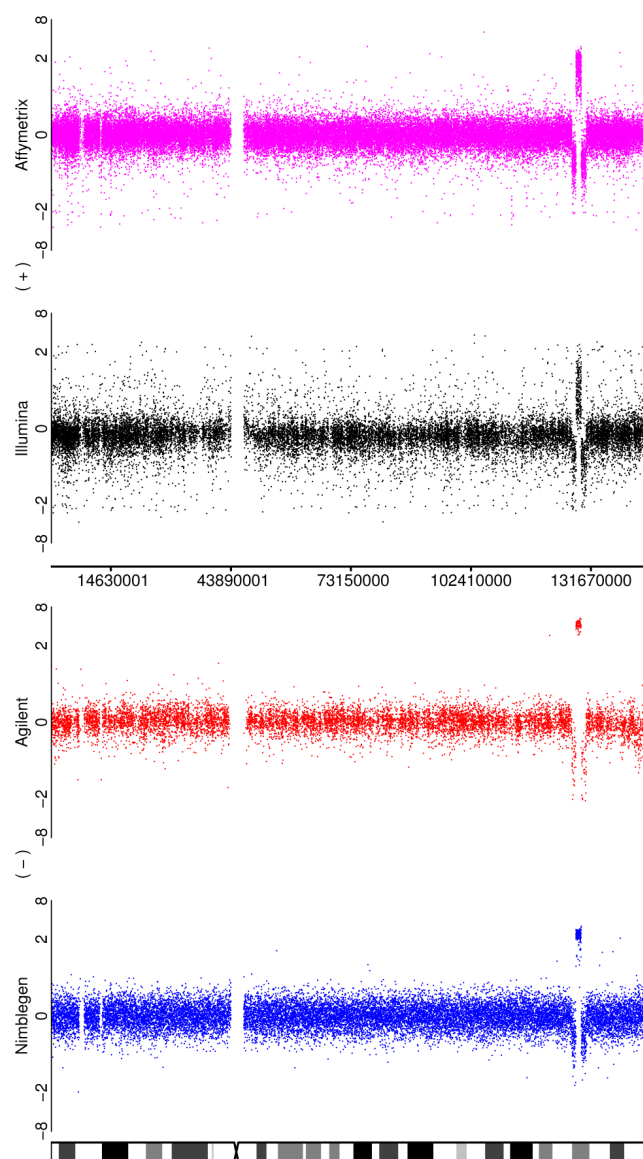


Figure 4
Illustrating the ability of the platforms to detect high amplitude focal amplifications and other subchromosomal events. Depicted are the log-ratios for a comparison of the SUM159 cell line to a pooled normal reference for chromosome 8. A deletion, duplication, deletion aberrations pattern is clearly visible for all four platforms in the region of $x = 130$ Mb.

Agilent and Affymetrix do generally best, but Nimblegen does a much better job of identifying the amplifications than it does for the neighbouring deletions.

iv.) Small regions of gain/loss as exemplified by copy number variation between normal HapMap individuals

A total of 79 sites of copy number variation have been identified between the two HapMap individuals assayed in this study using an older technology, namely a custom whole-genome tiling path array developed at the Wellcome Trust Sanger Institute [19]. These variants were validated across multiple hybridizations and also via PCR. For a full list of locations see Additional File 3. Examination with these higher resolution technologies suggests that some of the sites actually form one larger variant, but we shall treat them as separate sites for this analysis. Many of the sites showed no sign of variation with any of the platforms, and concordance amongst platforms was high. Due to the nature of these small changes, it is not uncommon for a platform simply to have no probes in the region of interest. This varies between platforms, with probe density being influential, but not the only factor.

The 79 sites were assessed by eye to see if they provided evidence of variation (the plots of all these regions are available in Additional File 4). Rating each CNV as clear, tentative, absent, or not covered, we summarize the results in Table 4. Naturally, there is an element of subjectivity in this type of assessment, but the overall picture is clear. Affymetrix and Agilent identify the greatest number of variants, but Agilent fails to cover a fair number (18 out of 79).

Notably, since some platforms (both Affymetrix and Illumina) have been designed to cover known CNVs and to target ‘unSNPable’ regions of the genome with copy-

number non-polymorphic probes, this rate will be misleading if one is interested in identifying novel CNVs. Nimblegen has more probes than Agilent, and a similar number to Illumina, but does not attempt to target known interesting regions with this version of the array. Thus Nimblegen may well do relatively better with novel sites. That said, the evidence here is that even if novel sites have coverage, the platform may struggle to identify them as CNVs. Illumina cover more of the regions than do Affymetrix, but do not provide the clarity of change over these small intervals.

Two CNV regions are shown in Figures 5 and 6. In Figure 5, CNV #58 is depicted and one can see that all of the platforms would identify it (with Affymetrix perhaps being the least clear). The ‘typical’ CNV #38 is depicted in Figure 6. Here, three of the platforms greatly reduce their coverage in the region of interest (Nimblegen being absent altogether), while Illumina exhibits good coverage. Despite this, the Affymetrix and Agilent probes that are in the region are quite clear, whilst Illumina is only convincing through weight of numbers.

Detection of characterized copy number aberrations

We address other measures of performance by making use of aberrations that have previously been reported to occur in the cell lines or have been broadly described to manifest in breast cancer. An examination of the six tumour samples (Table 5) reveals that there is little difference in the ability of the platforms to spot the large aberrations associated with cancer, with the exception that changes are harder to spot in Nimblegen than with the other platforms. The tumours themselves differ substantially, with T2704, T2706 and T2707 exhibiting far fewer aberrations, although we note that this may be a reflection of the sample’s cellularity. Figures 7, 8, and 9 highlight some of the aberrations observed in these tumours. For example, Figure 7 depicts Chromosome 17 for Tumour 7214 on all four platforms. Figure 8 reveals for Tumour 7207, the area surrounding the ADAM3A gene and Figure 9 the ERBB2 gene is shown. While none of the tumours here exhibits amplification of ERBB2, it is surprising to note how poorly represented this frequently amplified cancer gene is on the Illumina platform, although coverage is greater in the latest generation of the array.

Cellularity

The data set we present allows for the realistic comparison of platforms when considering copy-number changes in tumours. Tumour samples are often affected by stromal contamination [13] and to represent this, not only do we present 6 tumour samples of varying degrees of cellularity (see Additional File 5 for cellularity and clinical information for all samples), but a number

Table 4: CNVs observed between two HapMap samples

Platform	Clear	Tentative	Not Covered
Affymetrix	20	14	10
Agilent	19	16	18
Illumina	8	21	5
Nimblegen	9	13	14

Detection of reported germline CNVs between two HapMap samples across each platform, as adjudged from the plots in Additional File 4 by three analysts. Differences in probe density prohibited the blinding of analysts, but each platform was scored independent of the others. Indicated are the number of CNVs that are clearly apparent, the number for which there may be some evidence, which are labelled tentative, and those which are not covered. Naturally, there is a fourth category (not shown), which includes regions that are covered, but appear not to be CNVs.

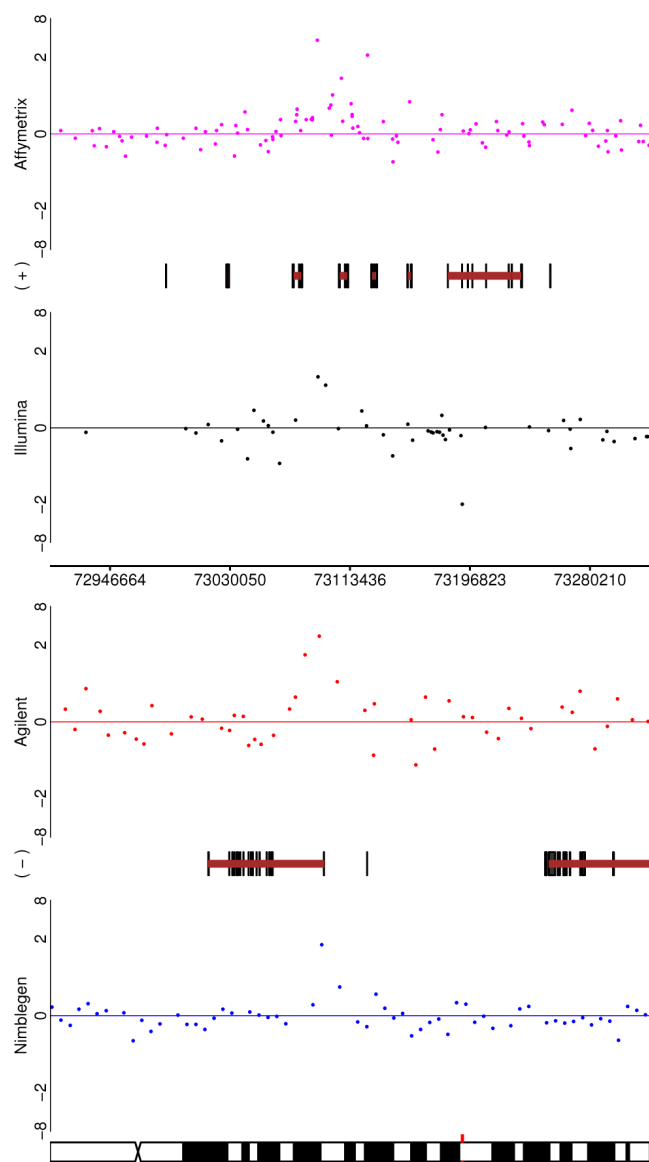


Figure 5
For one of the HapMap - HapMap CNVs (CNV58 from Additional File 3), depicted are the performances of all four platforms. The change is visible in each case, but with differing degrees of clarity.

of samples with simulated stromal contamination. Essentially, two of the tumour samples were diluted with their respective matched normal samples (7206: 30% tumour, 70% normal; 7207: 50% tumour, 50% normal) and two cancer cell lines were similarly treated (MT3 and SUM159: 30% tumour, 70% normal 7214).

We again consider the MT3 cell-line, this time in dilution, to see whether the anticipated copy number aberrations are visible (details of the expected copy

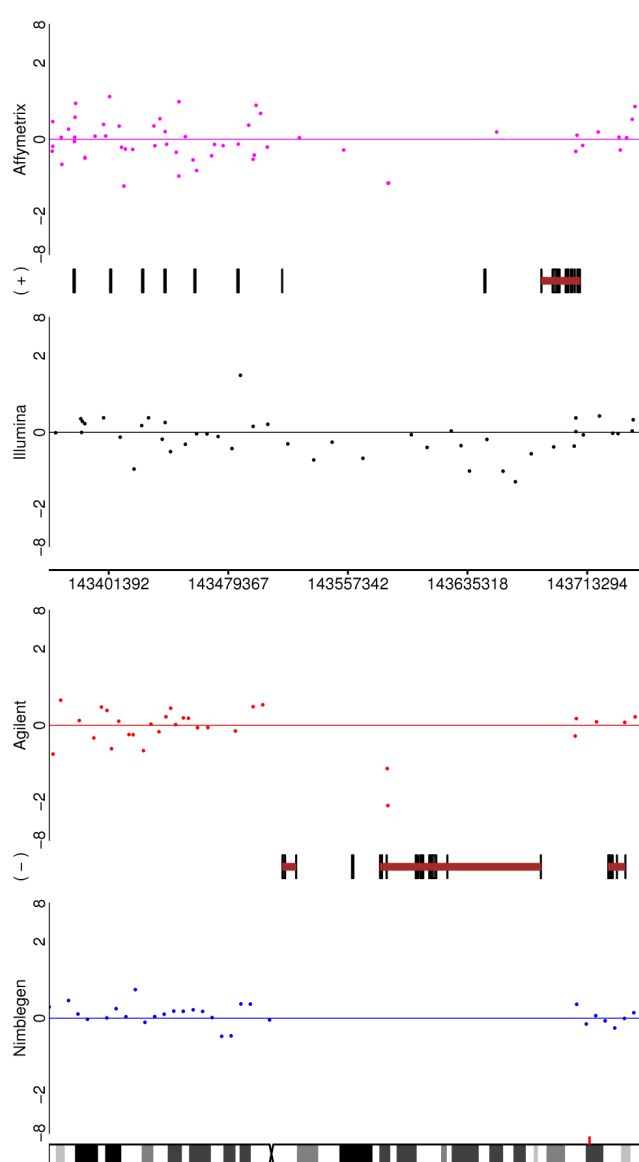


Figure 6
For one of the HapMap - HapMap CNVs (CNV38 from Additional File 3), depicted are the performances of all four platforms. In this case the variant is not obvious (or even apparent) in three of the platforms due to poor coverage of the region. Nimblegen has no coverage, and Agilent and Affymetrix have relatively low coverage. However, these last two platforms do show the copy number variation with what probes they have. Illumina is less convincing on a probe-by-probe basis, but successfully demonstrates the CNV through sheer number of probes in the region.

number alterations for the cell-lines are given in Additional File 6). In Figure 10, equivalent to Figure 2 but for simulated 70% stromal contamination, the

Table 5: Detection of anticipated aberrations across platforms for the 6 tumour samples

	T1975	T2701	T2704	T2706	T2707	T2714
Gain 8p	possibly 8q (all)	also 8q (all)	none	none	none	none, but gain on 8q (all but Nimb)
Gain 1q	all	all		all	none	all but Nimb
Loss 16q	partial (all)	and gain 16p (all)	none	and gain 16p (all)	none	none
Amp 8q24	all	all	none	none	none	all but Nimb
Amp 11q13	none	all	none	none	none	none
Amp 17q12	none	none	none	none	none	all
Amp 20q13	all	Affy and Agil	none	none	none	none
Del 13q14	all	all	none	none	none	none
Del 9p21	all but Nimb	all	none	none	none	none
Del 17p13	none	all	all	none	none	all

The detection of anticipated aberrations in each of the 6 tumour samples is reported for each platform.

benefits of direct competitive hybridization are seen. The two CGH platforms provide much clearer evidence of copy number differences between the chromosomes (as might be anticipated following previous studies [13]), and of the two, Agilent outperforms Nimblegen. It is not unreasonable that a direct comparison is better able to detect small changes such as those anticipated here. Note that as in Figure 2, no allowance has been made for probes targeting pseudoautosomal regions, which may explain the odd behaviour of the Y chromosome.

Figure 11 illustrates a zoomed-in region of chromosome 8q for a SUM159 dilution, similar to Figure 4 for the undiluted samples. As expected, all of the platforms exhibit some signal attenuation, but each is still able to detect the amplification. Notably, Agilent is clearly the least affected and in fact robustly detects the alteration at nearly the same level as the undiluted case, attesting to the sensitivity of this platform. In contrast, all of the platforms struggled to detect a moderate loss in the same sample.

Discussion

Discussion of results

The ability of a platform to detect a particular aberration is a function of the distribution of probes in that region and the reliability of those probes. Of the two SNP-based platforms, there is little difference in terms of quality of individual probes, but those on the Affymetrix arrays are more numerous. That said, Illumina's strategy for locating probes means that there are locations where this platform offers greater coverage (cf. the known CNVs and the MHC-similar region 5q13, consistent with Illumina's stated design intent which also sees a greater

focus on SNPs near RefSeq genes than does Affymetrix) but also some (such as the ERBB2 region) where they are lacking. The coverage of smaller features such as CNVs and genes is an important consideration in the choosing of a copy-number platform, as broadly speaking all of the platforms examined can identify large deletions and duplications.

A curiosity is that Illumina fails to identify robustly the chromosome 13 arm gain in the MT3 cell lines, suggesting an issue with the normalization applied by *BeadStudio*, but the main concern is the Nimblegen platform, which fails to spot some large aberrations in tumours T7195 and T7214. Of the two standard arrayCGH platforms, Agilent's performance is clearly superior. Not only is the Agilent data of high-enough quality to call aberrations from fewer probes than the other platforms, but also the ability of the Agilent platform to quantify aberrations appears to be superior. All of the platforms suffer from variation induced by probe design, related either to probe length, GC content or other aspects. Additionally, the quality of SNP and CGH probes on the Affymetrix and Illumina platforms may not be equivalent. Thus when choosing a platform one must consider not only the probe coverage in regions of interest, but also the quality of those probes.

Explanation of cellularity findings

The comparison for the diluted tumours is more complicated due to their pre-existing stromal contamination and the fact that aberrations of these tumours have not previously been well documented. Inspection of aberrations in the dilution of Tumour 7207 revealed one curiosity. Figure 12 depicts the area around the ADAM3A gene for

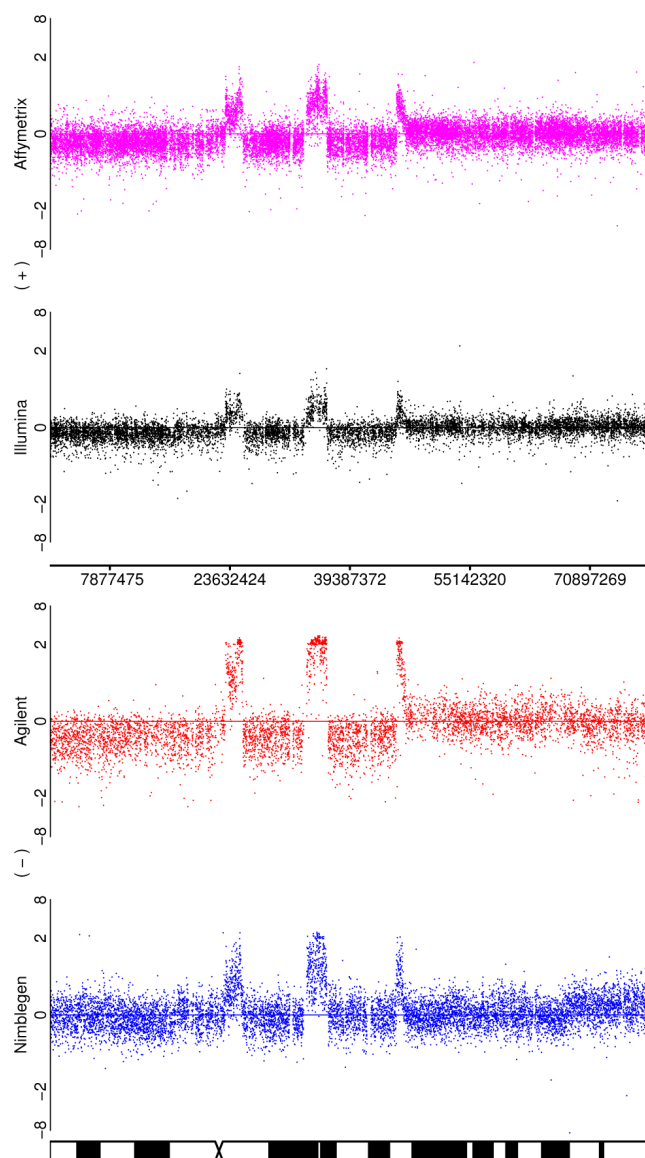


Figure 7
Depicting Tumour 7214, Chromosome 17 for the four platforms (genes not shown).

each of the four platforms as in Figure 8, but here for a dilution (50%). Surprisingly, we observe that all platforms more robustly detect the loss in the dilution hybridization. We note that this sample had the lowest cellularity (40%) of all the tumours assayed and that many common aberrations were not observed in this sample. This raises the possibility that the normal sample might actually represent a preplasia and brings into question the composition of the tumour. Indeed, subsequent expert histopathological examination of this sample revealed that the tumour section was likely comprised of inflammatory infiltrate rather than invasive

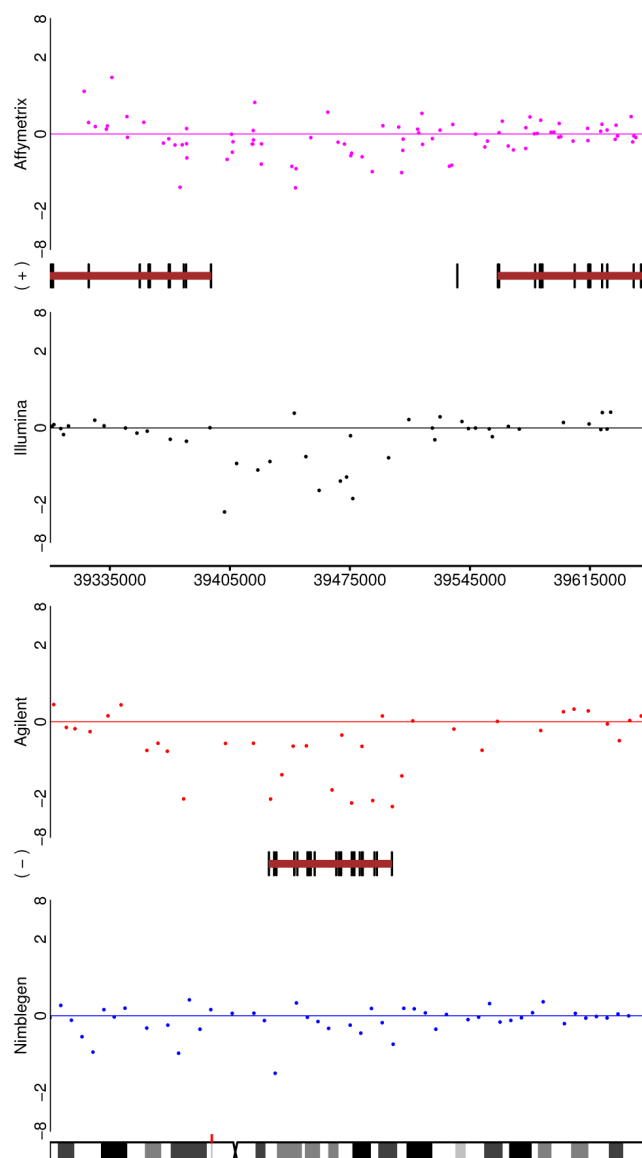


Figure 8
Depicting, for Tumour 7207, the area around the ADAM3A gene for the four platforms.

tumour cells. Upon examination of the matched normal sample for another tumour (7214) it was noted that the tissue contains substantial ductal carcinoma *in situ*. Despite the observation that this sample was used for dilution of the SUM159 cell line, this should not affect the previous discussion of the results since the 8q change was specific to that cell line and not observed in tumour 7214. It is noteworthy that examination of the array results prompted these findings, as this highlights the utility of microarray-based copy number assessment to detect subtleties in sample composition.

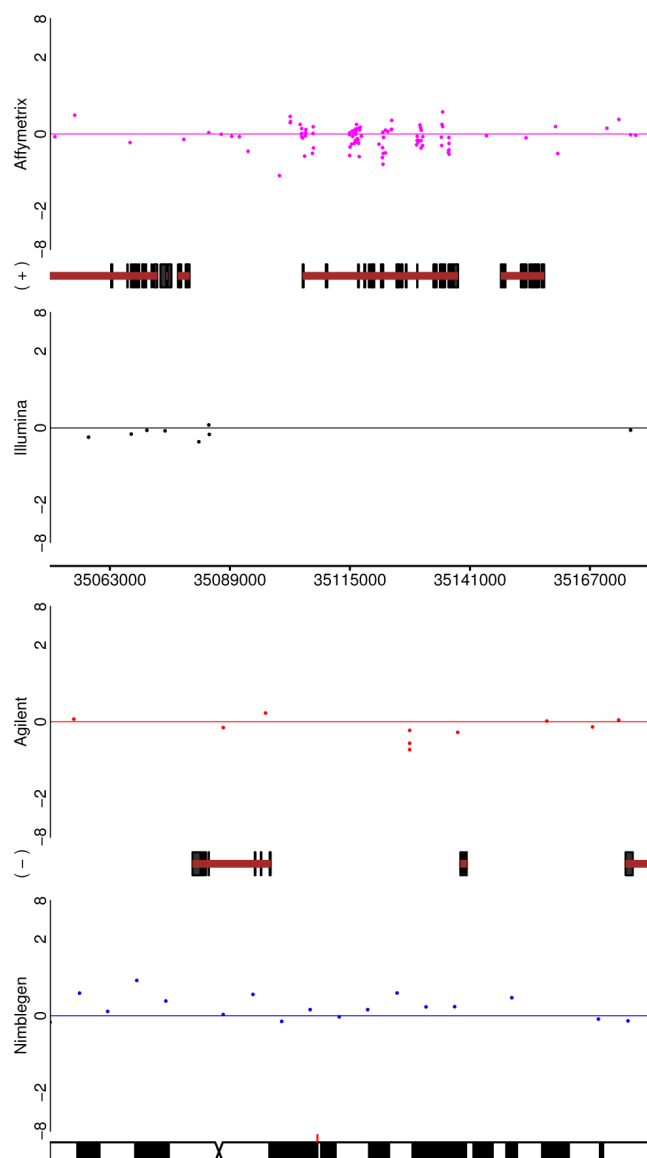


Figure 9
Depicting, for Tumour 7207, the area around the ERBB2 gene for the four platforms. Note the poor coverage of the Illumina platform.

Chemistry

In comparing microarray technologies, it is also important to keep in mind some of the more subtle differences between them in terms of the protocols, chemistry, and detection methods. For example, while both Affymetrix and Illumina are SNP-based copy number profiling platforms, there are important differences in their chemistries. The Affymetrix GenomeWide SNP 5.0 whole genome genotyping assay (as well as the newer SNP 6.0 array and older generations of this platform, namely the 10 K - 500 K arrays) all employ a complexity reduction procedure

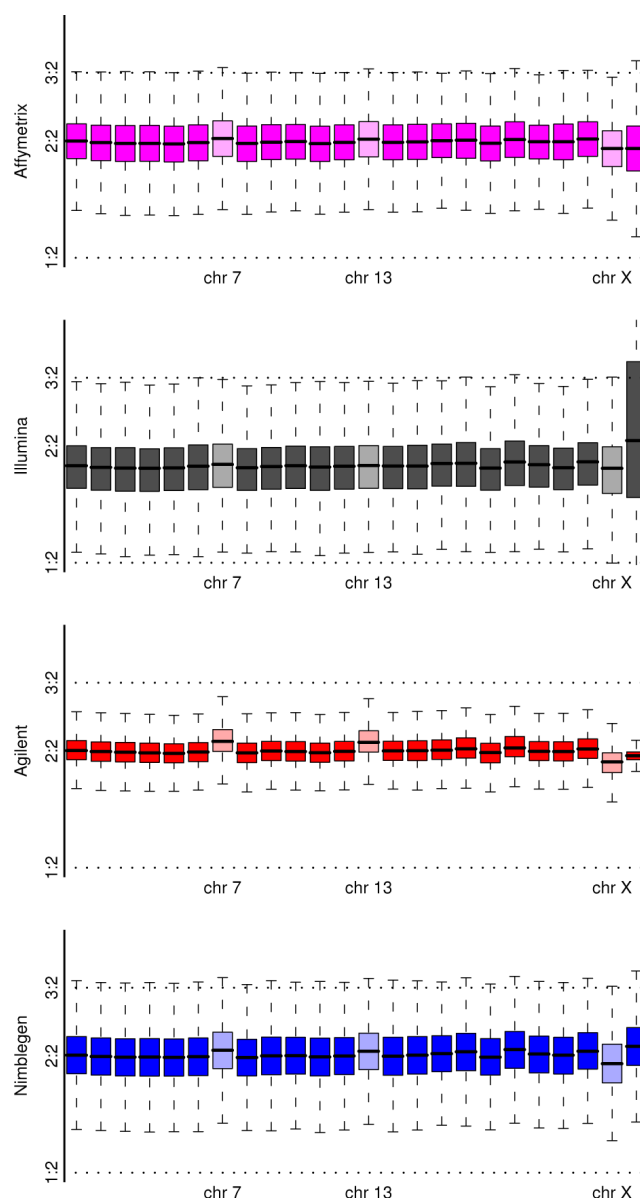


Figure 10
For comparison with figure 2: Depicting, for a dilution of the MT3 cell-line, compared to a pooled normal reference, a boxplot of the log-ratios from each platform broken down by chromosome. Also indicated are theoretical markers for a single copy gain and a single copy loss at this dilution level. The three chromosomes with known aberrant copy number are indicated.

similar to that first described for representational oligonucleotide microarray analysis (ROMA) [20] in order to increase the signal-to-noise ratio. Essentially, the DNA is digested with the restriction enzymes *NspI* and *StyI*, ligated to adaptors that recognize the cohesive four base-pair overhangs, and amplified using a universal primer that

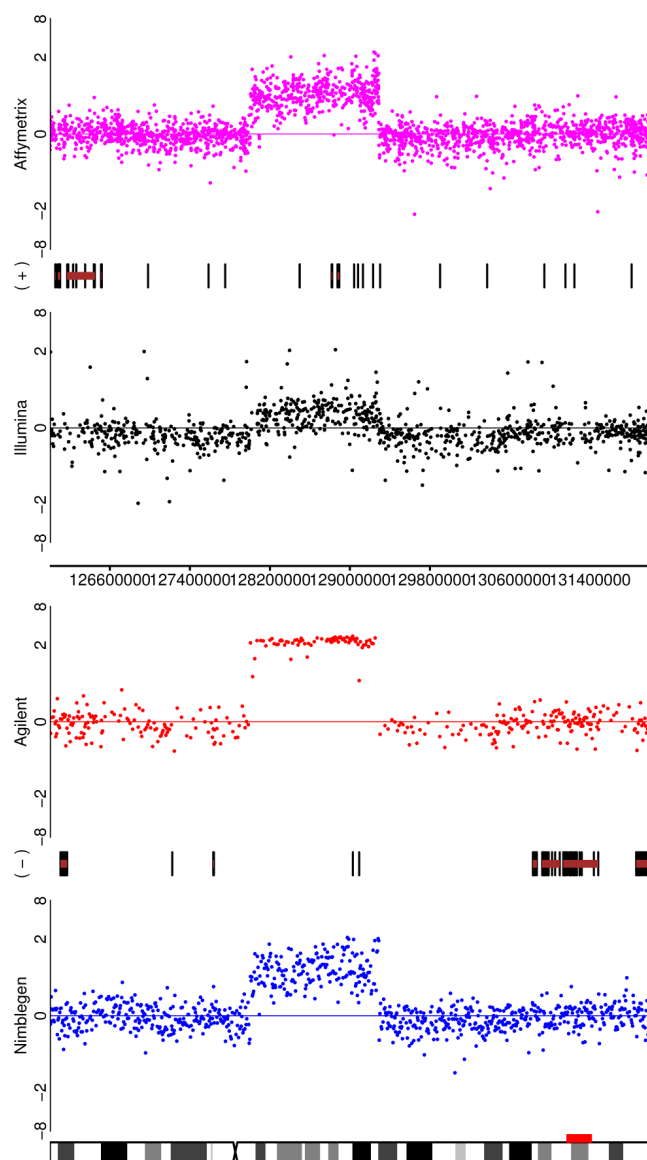


Figure 11
Depicting, for a dilution of SUM159, the 8q region for the four platforms.

recognizes the adaptor sequence. The amplified DNA is subsequently fragmented, labelled, and hybridized to the oligonucleotide array. While the amplification of only the smaller restriction fragments improves the signal-to-noise ratio, these values still remain below that observed for BAC arrays, and the complexity reduction can potentially lead to the differential representation of certain genome regions and hence false positives. Also, since individuals vary in their restriction digestion profiles, certain probe ratio values may depend on differences in restriction fragment size rather than actual copy number variation [21].

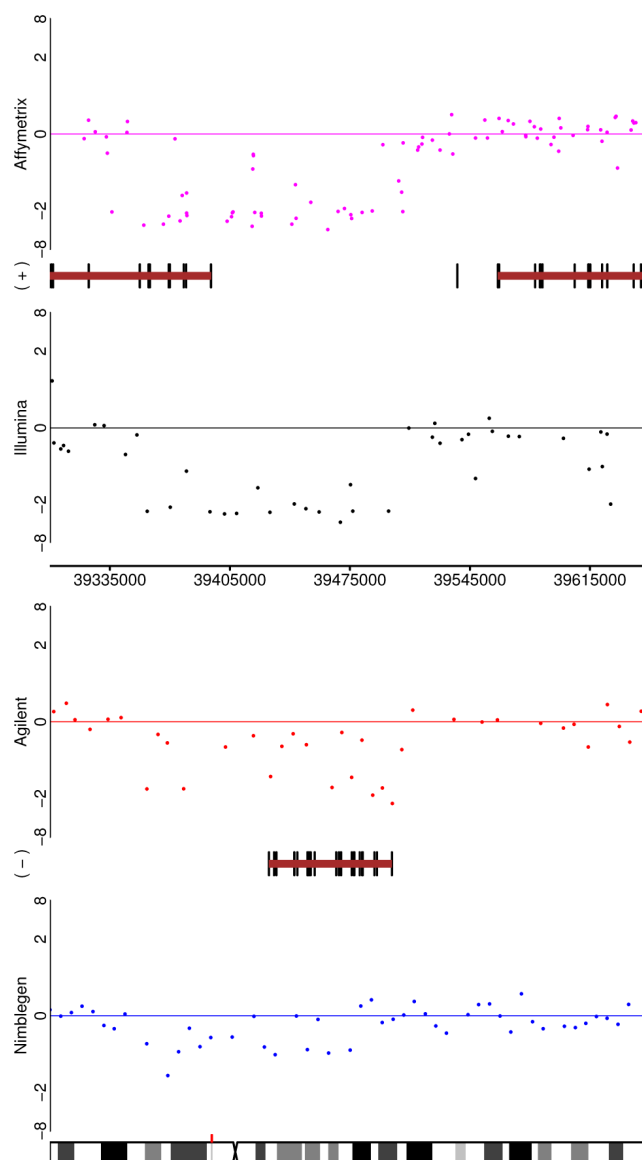


Figure 12
Depicting, for a dilution of Tumour 7207, the area around the ADAM3A gene for the four platforms.

In contrast, the Illumina whole genome genotyping protocol for the 370 HapMap Duo bead array (Infinium II technology) involves an isothermal genome amplification step (non-PCR based), fragmentation, hybridization to an oligonucleotide bead array, SNP detection based on a single-base extension reaction (SBE) on a single bead type with differentially-labelled terminators, and signal amplification. Thus the detection step, for the Illumina Infinium II assay is based on an enzymatic discrimination step (SBE for Infinium II, allele-specific extensions for Infinium I) rather than by hybridization as for Affymetrix. Illumina claims that

the isothermal amplification step does not result in the preferential amplification of one allele [22].

New Platforms

All of the manufacturers now offer products with more features than those compared in this report: the Affymetrix GenomeWide SNP 6.0 array, the Illumina 1 M-Duo array, the Nimblegen Ultra-High Density CGH array with 2.1 million features, and the Agilent Human CGH 1×1 M array. All but the Affymetrix chip come available with fewer features but multiple arrays on the chip (the Illumina platform starting with 2 arrays on the chip); the ability to run multiple samples in parallel is of great potential value for sensitive experiments. Also worthy of note is that the Nimblegen and Agilent platforms offer full customization of content, while Illumina offer limited customization.

As the coverage of platforms increases, many of the subtleties that we have observed will have decreased impact on the conclusions. The Illumina coverage of ERBB2, for example, is satisfactory in the latest generation of chip. It remains to be seen whether the manufacturers have been able to maintain probe quality in the next generation of products. We have already commented that the second generation of the Nimblegen platform featured here has seen a revision of probes to improve performance.

The other disappointing performance we have witnessed was that of the Affymetrix SNP5 platform for the Y chromosome. The newer Affymetrix SNP6 platform contains nearly 10 times as many Y chromosome probes, including approximately 900 SNP probes (recall that SNP5 contained only non-polymorphic probes for the Y chromosome). Of the 997 SNP5 Y chromosome probes, 127 (12.74%) are retained on SNP6. Hence SNP5 probes makeup only 1.34% of the total SNP6 Y chromosome repertoire. Using a publicly available Affymetrix SNP6 HapMap X chromosome titration data set [23], we compare the sensitivity and specificity of the SNP5 and SNP6 platforms in Figure 13. The SNP6 platform performs similarly to SNP5 in the detection of a 2:1 copy number alteration, whereas for a 1:0 alteration the improvement is striking.

Alternative analysis methods

It should be noted that we have made use of manufacturer-provided tools, where available, for pre-processing this dataset. This was intentional, as the choice of optimal tools is platform-specific, especially since older platforms will likely benefit from more mature analysis tools. For example, we employed the *BeadStudio* software to summarize the Illumina data as this is manufacturer-supplied. Likewise, Nimblegen supplied NimbleScan pre-processed data. In contrast, Affymetrix do not offer comprehensive support for copy

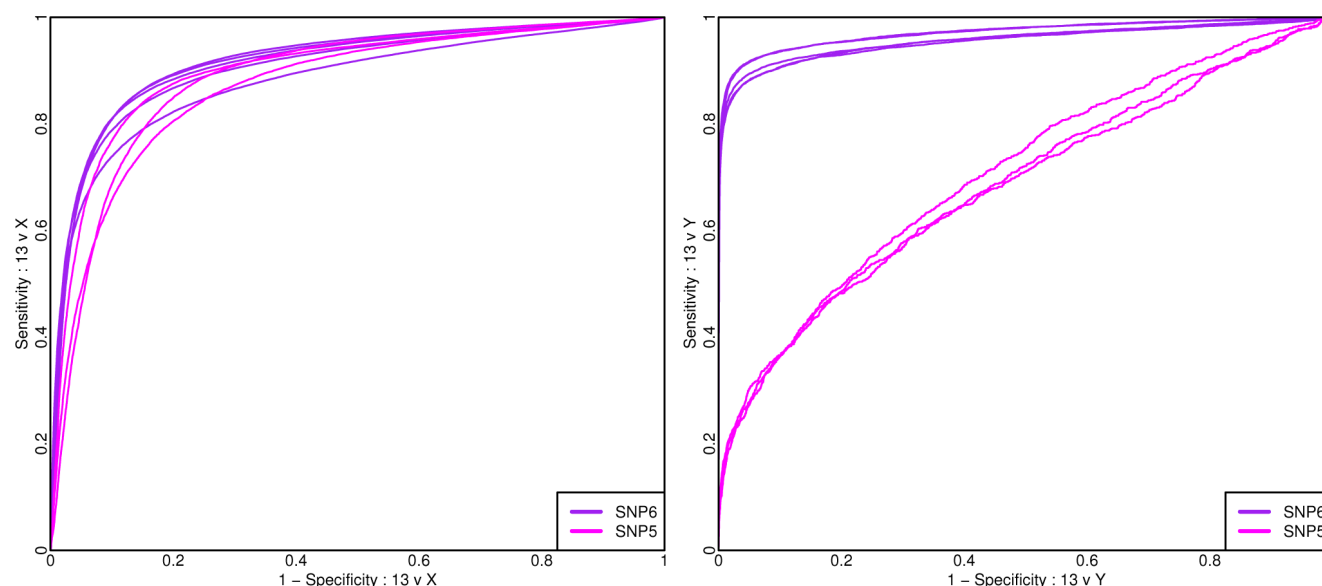


Figure 13

For comparison with Figure 3. Here a comparison of the Affymetrix SNP5 and SNP6 platforms are shown. ROC curves are presented to assess the performance of a single probe/probe-set for distinguishing the log-ratios associated with differing copy numbers from the log-ratios of chromosome 13 (where copy-numbers should agree) for the HapMap pair of samples. For SNP6 five replicate HapMap/HapMap (NA15510 vs NA10851) comparisons are shown using raw data available from the Affymetrix X chromosome titration study.

number analysis of the GenomeWide SNP5.0 platform so we employed the open source *aroma.affymetrix* [24] software, one of the few tools for pre-processing this relatively new SNP-CGH platform. Similarly, as Agilent do not provide free software for pre-processing of their aCGH data, standard open source methods were employed for this well-established platform. Although beyond the scope of this study, it is of interest to compare alternate pre-processing (and segmentation) methods for each of these platforms as this could influence the results obtained. In particular, the consideration of Illumina SNP data at the bead level could yield considerable improvements since this would enable calculation of the within-bead-type correlation or covariance [17] as well as more detailed quality assessment. For example, we were able to identify spatial artefacts in the Illumina data in this study (Additional File 7) that would benefit from the BASH tool [25] implemented in *beadarray* [26] although this method has not yet been fully implemented for Illumina genotyping data. Additionally, there were some issues with small, localized failures of image registration that could only be addressed by bead-level pre-processing and that would undoubtedly improve the quality of the Illumina results if addressed successfully.

Conclusion

It is important to stress that there is no straightforward way to compare fairly copy number profiling platforms in a general manner. As such, the results presented here describe the detection and qualitative comparison of raw copy number alterations across four platforms in tumour samples for which both matched and pooled normal DNA were available and in two established cell-lines. Copy number variation in normal HapMap individuals was also compared using the same platforms. Whilst we have sought to avoid analytical techniques that are objective, but that we deem undesirable for the stated reasons, we have focused on graphical comparisons that are, of course, prone to subjectivity. In any case, the competing platforms have different merits, and users need to make subjective decisions based on their individual requirements.

Although there are substantial differences in the design, density, and replicate structure of the probes, the comparison indicates a generally high level of concordance between platforms. As expected, all platforms were able to detect large aberrations in a robust manner. However, some focal amplifications and deletions were only detected on a subset of the platforms. In particular, Nimblegen failed to detect numerous aberrations that were clear in the other platforms even when probes were tiled in the region of interest. This finding is perhaps not

surprising given that this platform exhibits 2-4 fold greater variance amongst replicate probes and variances an order of magnitude greater for replicate array comparisons. In general, for the aCGH-based platform Agilent was the best performer and for the SNP-CGH platform, Affymetrix tended to outperform Illumina. An added bonus is that both Affymetrix and Agilent require only 0.5 µg DNA as starting material, thus removing this consideration from the platform decision. Another potential consideration is the quality or source of DNA (e.g. the use of paraffin-embedded samples [13]), for which some platforms may be more forgiving.

Our study differs from previously published ones in that we employ primary breast tumour samples rather than cell-lines. As noted previously, this introduces additional complexity due to the possibility of stromal contamination [13]. Further to this, we have also made use of cell-line dilutions and well-characterized HapMap samples to evaluate copy number alterations across platforms. That we also conclude that Agilent performs best on a single-probe comparison is of interest because we are comparing newer platforms, yet we must keep in mind that the performance of platforms from generation to generation cannot be assumed to be constant.

In the new generation of arrays, Agilent have addressed their primary weakness by increasing probe coverage. Similarly, Nimblegen have modified their probe design in order to improve performance. Both Affymetrix and Illumina have increased probe coverage with Affymetrix introducing slight modifications to probe design. If Agilent have maintained probe quality, it seems likely they will remain the leader, but Nimblegen may close the observed gap. For the SNP-CGH arrays, it seems likely that Affymetrix will continue to perform well. The availability of data from these new platforms will enable comparisons with previous generations of arrays for the purposes of meta-analyses and the like.

Obtaining reproducible, high-resolution copy number data with high sensitivity and few false positives is the gold-standard objective for any such study. However, there are always tradeoffs and a critical assessment of the goals of the project and underpinning biological questions can help select the most suitable platform. For example, breakpoint precision, which is dependent on the local resolution, is likely more critical for mapping novel tumour suppressor genes and oncogenes, than for a more general survey of aberrations where little follow-up validation is planned. Additional considerations that might influence the choice of platform include probe coverage (whether gene-centric or uniformly spaced, targeting non-coding elements) and the ability to assay genotypic information, and hence allele-specific copy

number and copy neutral loss of heterozygosity. If matched normal samples are available, it might be advantageous to exploit the direct comparison design offered by dual-channel technologies. In large-scale studies, it may also be useful to validate the higher-density SNP-CGH findings using a subset of samples on a lower-density, but more sensitive, platform. The results described here provide a guide for platform selection and study design, and the dataset a resource for more tailored comparisons.

Methods

Study design

The state of the art in terms of commercially available platforms for genome-wide CNA is constantly evolving. Here, four leading platforms were compared: the Affymetrix Genome-Wide Human SNP Array 5.0, the Agilent High-Density CGH Human 244A array, the Illumina HumanCNV370-Duo DNA Analysis BeadChip, and the Nimblegen 385 K oligonucleotide array. Several important differences exist between these platforms. Beyond the fact that the Affymetrix and Illumina employ a single-channel hybridization scheme, whereas Agilent and Nimblegen use a dual-channel competitive hybridization protocol, the former are also SNP-CGH platforms, while the latter are not. Other differences in the design of these platforms include the probe-length and probe-density. Whereas Nimblegen employs 45-mer to 85-mer probes, Agilent 60-mer probes and Illumina 50-mer probes, Affymetrix probes are considerably shorter at 25 nucleotides. In terms of probe-density, the Affymetrix SNP 5.0 array contains 500,568 SNP probes and an additional 420,000 non-polymorphic probes to facilitate studies of germline copy number variation in association studies. The Agilent 244A array contains computationally pre-selected probes that have been experimentally optimized for genomic hybridization with a bias towards gene-rich regions. The Illumina CNV370 array includes 318,000 SNP markers plus 52,000 markers targeting 14,000 additional CNV regions. Lastly, the Nimblegen 385 K array contains 386,165 isothermal oligonucleotide probes with relatively uniform genome coverage. Due to resource availability, two of the platforms (Agilent and Illumina) were processed in-house, whereas for the other platforms the samples were hybridized at a commercial vendor (Affymetrix and Nimblegen).

Sample Choice

Two representative cell lines (MT3 and SUM159) were selected based on the presence of known chromosomal aberrations so as to provide markers of a platform's performance. The MT3 colorectal cell line contains a single copy gain of chromosome 7 and isochromosome 13 [27,28]. The SUM159 breast carcinoma cell line is

also reported to have several notable changes including a loss on chromosome 5q and gain on chromosome 8q24 [27,28]. The ability of the various platforms to detect known focal amplifications was assessed using a panel of six tumour samples. To assess the effect of using a matched normal as compared to a pooled normal as the reference against tumour samples, a single replicate was included for each matched normal sample. Additionally, to ascertain the effect of cellular heterogeneity due to stromal contamination in detecting CNA, several dilution experiments were included for the two cell lines and two of the tumours such that a mixture of either 30% cell line (tumour) with 70% normal or a 1:1 ratio was hybridized to the arrays.

Two 'normal' samples (NA15510 and NA10851) from the Human HapMap study [29] were also selected to assess the detection of naturally occurring regions of copy number variation, as they have been characterized extensively [30-33] and are recommended for use as a standard control in all studies [21]. Further, they provide an example in which gross abnormalities are not expected. Moreover, sample NA10851 is male, allowing for a controlled assessment of the platforms performances by examination of the sex chromosomes in the HapMap comparisons.

Each sample was hybridized to the single-channel platforms in triplicate, with the exception of the pooled normal samples, which were performed in duplicate. For the dual-channel platforms, tumours and cell-lines were hybridized against pooled normal tissue in duplicate, and the tumours were additionally hybridized against matched normal tissue. The HapMap samples were hybridized against each other in duplicate, as was a pool vs. pool hybridization. Additionally dye-swap hybridizations were performed for the HapMap samples and the MT3 cell-line. In all platforms, save for Nimblegen, some hybridizations were discarded under quality control procedures. Nimblegen only returned data for hybridizations that satisfied their quality control criteria.

Patient material and cell lines

Samples were collected in the year 2000 at Addenbrooke's Hospital, Cambridge, UK from female patients ranging from 41 to 83 years old. These samples correspond to fresh frozen biopsies and surgical resection samples and the resultant fresh breast tissue was stored in the Addenbrooke's Hospital tumour bank. Ethical consent was obtained for all patient samples. The MT3 cell line (with a single X chromosome, suggesting male origin) was obtained from its originators [34], and has been shown to be identical to the colorectal cancer

cell line LS174T based on SNP analysis [35]. This cell line exhibits an almost normal karyotype, apart from trisomy 7 and isochromosome 13. The SUM159 breast carcinoma cell line was obtained from the originators [27]. SUM159 is a hyperdiploid cell line with a modal chromosome number of 47 and nine structural translocations.

All human samples used for this analysis were obtained with informed consent from patients and the study was performed with appropriate REC and NHS R&D approval.

DNA extraction and purification

Tumour DNA was extracted from 25 × 10 µm sections manually using the DNAeasy kit (Qiagen, Valencia, CA). Matched normal DNA was obtained by homogenizing tissue in 180 µl of ATL buffer with Precellys, followed by extraction with the DNAeasy kit (Qiagen). For the cell lines, DNA was extracted using the proteinase-SDS method [36].

Array hybridization and analysis

Affymetrix

Genotyping using Affymetrix Genome-Wide SNP 5.0 arrays was performed according to standard Affymetrix protocol (at AROS, Denmark) using 0.5 µg DNA. Log₂ signal intensities were measured from the raw data derived from the scanned image. Signal intensities were corrected for allelic crosstalk and offset for SNP probes and for offset for copy number non-polymorphic probes (CN probes), and probe signals were rescaled so that all probes (excluding those on the X & Y chromosomes) have the same average across arrays [24]. Probe-level data were summarized, wherein probe signals for SNP probes were averaged across replicates and summarized between alleles; probe signals for CN probes were unchanged since they are generally-unreplicated single-probe units. Signal intensities were shifted by 300 units to avoid negative signals that might result following calibration for allelic crosstalk and due to random errors around zero. Fragment-level normalization was then performed to correct for systematic differences in the amplification efficiency of PCR on fragments of varying length and deviations from the 50/50 NspI/StyI mixture. This procedure is a multi-chip method, which estimates the baseline effects as effects observed in a robust average across all arrays and hence should cause systematic effects across arrays to cancel out. Raw total copy number estimates (on the log₂ scale) were obtained by comparing the summarized and normalized intensity values for a given cell line or tumour sample to the corresponding intensities from the reference array. Although 920,928 SNP probes and non-polymorphic copy number probes are present on the array, due to incomplete information

concerning a subset of the probes, 828,737 are analysed in this study. As noted above, Affymetrix data were corrected for fragment length effects as it has been noted that fragment length influences probe intensity, as does GC content [14]. Further, as for gene expression arrays, the sequence effect is position-dependent for Affymetrix SNP chips and importantly, fairly large differences in intensities can be observed for the different alleles as a result of sequence alone [37]. The effects of GC content are illustrated in Figure 14. The Affymetrix dataset consists of 50 arrays, as detailed in Additional File 8.

Agilent

The Agilent platform used is the Agilent Human Genome CGH Microarray Kit 244A. This platform uses just under 240,000 unique 60-mer oligonucleotide probes across the genome, with tighter coverage in the region of RefSeq genes, and claims to emphasize other interesting genomic features (miRNAs, promoters, etc) also. Experiments were performed in-house using 0.5 µg of DNA and either the Agilent labelling kit or the Enzo labeling kit. After hybridization and washing, the slides were scanned on an Agilent Microarray Scanner and captured images were analysed with Feature Extraction Software v 9.1.

Arrays were considered for analysis using a guideline DLRS threshold of 0.3. This is higher than the threshold

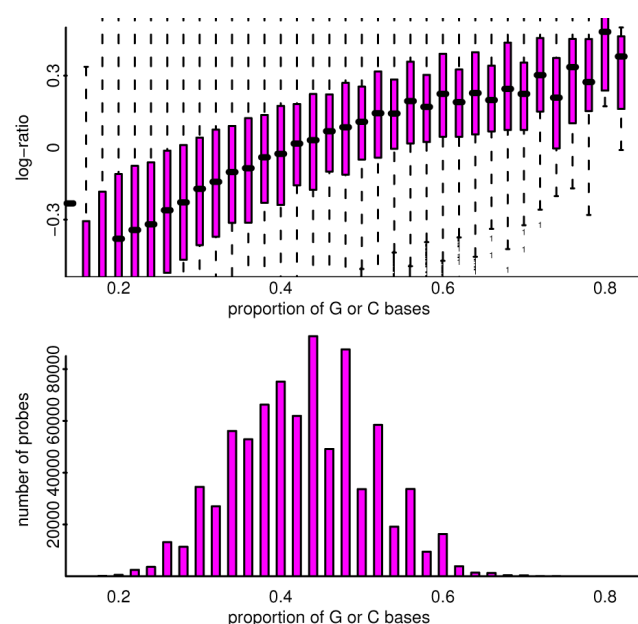


Figure 14
For a pool-pool log-ratio comparison from the Affymetrix platform, depicted are the effect and distribution of probe GC content. Top: Showing the effect of GC content on log-ratio. Bottom: Showing the distribution of probe GC content.

advised by Agilent, but that threshold does not allow for the large aberrations associated with tumour samples that will inevitably inflate this score. Where necessary (if multiple repeat hybridizations for a sample failed to bring the score down), hybridizations with a higher score were used to fill in gaps in the experimental design if they were judged to be acceptable. Similarly, some samples were not used despite passing the threshold if they were clearly problematic from a visual inspection. This resulted in 40 arrays remaining in the study (see Additional File 9). The Enzo protocol used for Agilent saw generally lower scores for this quality control measure, but saw an increase in the influence of probe length on the results from the array.

Based on the annotation information included in the Agilent output files, only 215,002 out of 238,162 (90%) of Agilent probes appear to be targeting 60 mer sequences (Figure 15), with the rest being shorter (as short as 45 mers in some cases). There is a marked relationship between observed intensities and target sequence lengths for the platform, with the probes targeting longer sequences generally generating lower intensities. This feeds through to having greater variance of log-ratio for the longer sequences. The effects are often more marked than in the example shown, and as a result

the non-60 mer targets have been dropped from the analysis. Intensities were background corrected via the *normexp* function in the *limma* package [38,39] and loess normalized to return log-ratios. No between-array normalization was performed; where between-array comparisons are made, we specify the steps taken to scale the arrays in question.

Illumina

Genotyping using Illumina CNV370-duo arrays was performed in-house according to the standard protocol with 0.75 µg DNA. Log₂ signal intensities were obtained using Illumina's *BeadStudio* software (ver.3). Following averaging of the per-allele replicates (16 on average), the A and B alleles are summarized, scaled and rotated to reduce allelic crosstalk on a per-array basis. Within *BeadStudio*, a paired analysis was performed for all contrasts of interest. The resultant log₂ ratios were then exported from the *BeadStudio* software to facilitate comparisons between platforms. Since the log-ratios were not centred around zero for the tumour samples relative to a pooled normal (while this was the case for the tumour samples relative to the matched normals), both subsets of assays were normalized under the assumption that median copy number is 2 and the median log₂ ratio is zero. The effects of the GC proportion on log-ratios are shown in Figure 16.

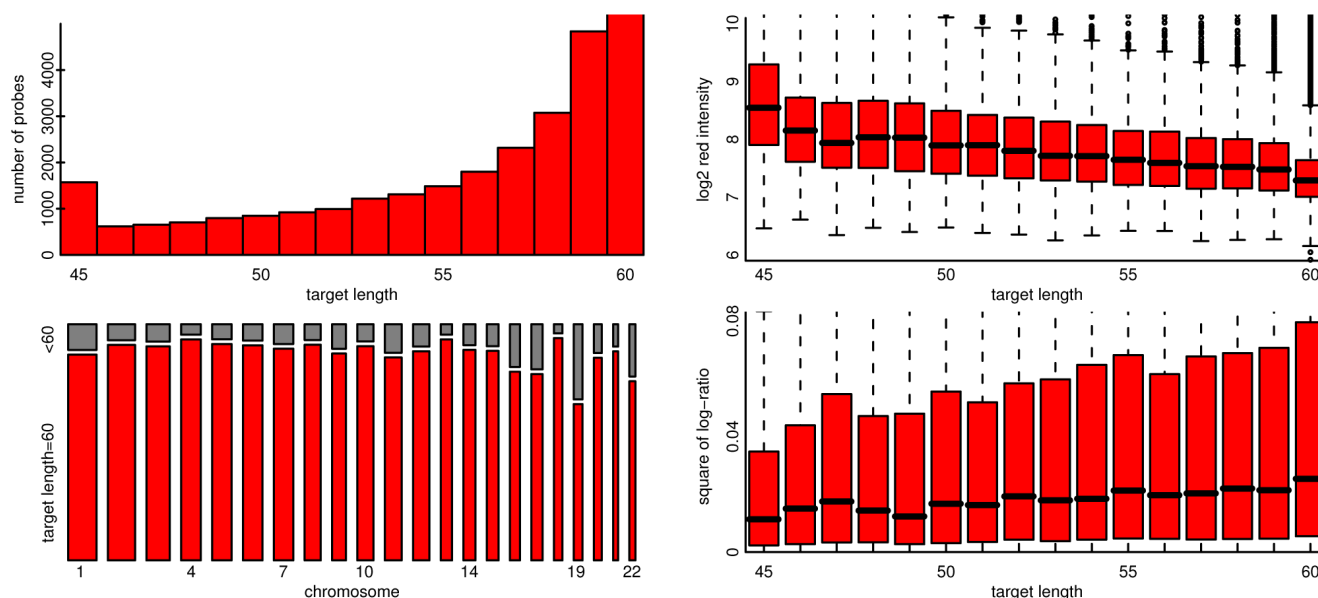


Figure 15

For a pool-pool hybridization from the Agilent platform depicted are the distribution and effect of probe target length. Top left: depicted are numbers of probes apparently targeting sequences of different lengths, with modes at 60 and 45. Bottom left: Shown are the proportions of probes, for each autosomal chromosome, that have target length 60; a proportion that is lowest for chromosome 19. Note that the width of the bar is proportional to the total number of probes on that array. Right: Two boxplots depict the associations between probe target length and intensity, and probe target length and log-ratio. 60 mer target sequence lengths are associated with lower intensities and greater variance of log-ratio.

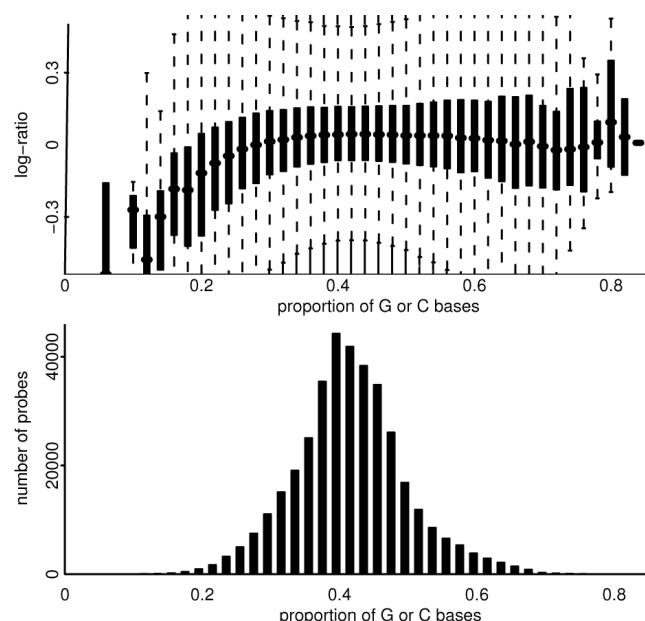


Figure 16
For a pool-pool log-ratio comparison from the Illumina platform, depicted are the effect and distribution of probe GC content. Top: Showing the effect of GC content on log-ratio. Bottom: Showing the distribution of probe GC content.

The Illumina dataset consists of 48 arrays, as detailed in Additional File 10.

Nimblegen

The Nimblegen platform used here is the HG18 CGH 385 K WG Tiling v1.0 array. This platform makes use of 385,000 oligonucleotide probes of length 50 mer to 75 mer. These probes are spaced along the genome with reasonable uniformity, unlike for the v2.0 array that followed, where probe locations were subject to more involved design. The experiments were performed by Nimblegen according to their standard protocol using 2.5 µg DNA, and were analysed using the processed and normalized data files supplied by Nimblegen. Nimblegen also report the lengths of the individual probes and the proportion of bases that are either G or C. The effects of the GC proportion are shown in Figure 17; there is a strong association between probe length and GC content, but still some evidence that probe length is influential even after GC content is considered (not shown). The Nimblegen dataset consists of 44 arrays, the details of which are in Additional File 11.

All analysis was performed in the R statistical programming language [40]. The arrays described in this study have been deposited in the Gene Expression Omnibus [41] with accession number GSE16400. Plots of each

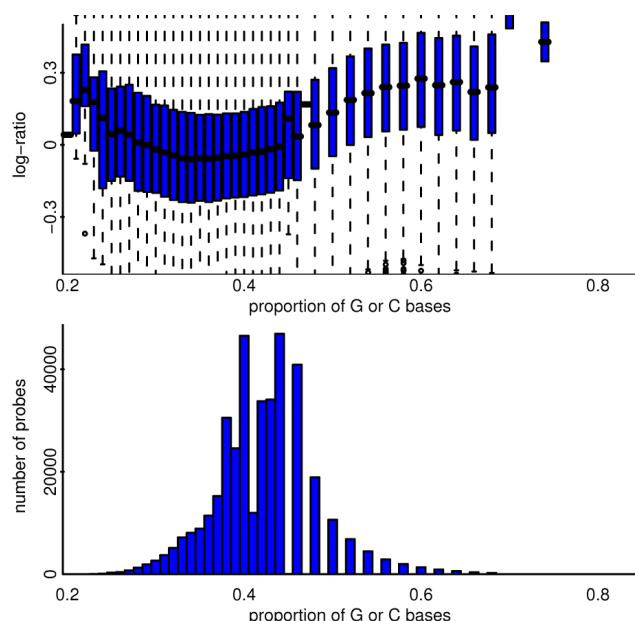


Figure 17
For a pool-pool hybridization from the Nimblegen platform depicted are the effect and distribution of probe GC content. Top: Showing the effect of GC content on log-ratio. Bottom: Showing the distribution of probe GC content. The median GC content is 0.42 (IQR 0.38 to 0.44), but is noticeably lower for chromosomes 4 and 13, and noticeably higher for chromosomes 19 and 22. Naturally there is a high spatial autocorrelation of probe GC content along the genome.

chromosome for each sample and platform are available in Additional File 12 and 13.

Plotting conventions

Where we have plotted relative copy number (log-ratio) against genomic location, we have used the best quality example for each platform. This may be the cause of a slight bias, as different platforms may have different numbers of replicates from which to choose, but since we are looking to establish the potential of the platforms, it is the appropriate approach. Replicates have not been averaged, as between-array standardization has not been performed, save for the case of CNV comparisons, where three replicates of each platform are comparable without standardization and the improved signal-to-noise allows for acceptable clarity with so few probes. Genomic location was taken from the supplied annotations for Agilent and Nimblegen, and likewise for Affymetrix and Illumina. For the different platforms the genomic location represents different properties (probe start, SNP location etc). However, on the scale on which we are plotting, this does not affect interpretation.

The scale for the y-axis for the plots is linear from -2 to 2, and linear also outside this region, but at a different rate. Most values lie in the -2 to 2 range and this needs to be our focus, but it is also important that we can depict more extreme cases. The discontinuity in the first derivative of the scale allows us to achieve this. As well as the log-ratios for the four platforms, we depict genes lying on the plus and minus strands, and a guide to the section of chromosome being illustrated. The information for these additional items was obtained from the *GenomeGraphs* [42] package in Bioconductor.

Where CNV locations are plotted, the nominal location lies within the middle two-fifths of the x-axis, allowing for easy use of the provided axis coordinates to identify that region. Throughout the paper, we adopt a convention of colour-coding for platforms: Affymetrix are represented by magenta, Agilent by red, Illumina by black, and Nimblegen by blue.

Competing interests

CNC owns Illumina shares.

Authors' contributions

CNC co-conceived the analysis strategy, analysed the Affymetrix and Illumina data, and drafted the manuscript. AGL co-conceived the analysis strategy, analysed the Agilent and Nimblegen data, and drafted the manuscript. MJD helped process and analyse the Illumina data. IS participated in the study design and prepared samples. JCM helped design the study and provided CNV expertise. JH supervised the Agilent and Illumina array hybridizations. SFC supervised the sample preparation. JB participated in the study design. ST contributed to manuscript preparation. CC conceived the study, participated in its design, and contributed to manuscript preparation. All authors have read and approved the manuscript.

Additional material

Additional file 1

Detailed description of platform features. An extension of Table 1. In order the columns represent i) The numbers and percentages of features by chromosome for each of the four technologies (cols B-I), ii) Within chromosomes, the numbers and percentages of features within the p arm (cols J-Q), iii) the core region in the p arm covered by all four platforms (cols R, S), iv) for each technology, in that core region, the probe density, the number of extra probes towards the telomere, the extra distance covered towards the telomere, the probe density in the extra region towards the telomere, the number of extra probes towards the centromere, the extra distance covered towards the centromere, and the probe density in the extra region towards the centromere (Affymetrix cols T-Z, Agilent cols AA-AG, Illumina cols AH-AM, Nimblegen cols AN-AU), v) as per ii) to iv), but for the q arm.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S1.CSV\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S1.CSV)

Additional file 2

Details of SUM159. Plots detailing the loss-gain-loss aberration on chromosome 8 of SUM159.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S2.PDF\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S2.PDF)

Additional file 3

A list of validated CNV sites for the HapMap/HapMap comparison. The full list of the 79 sites of copy number difference between HapMap samples NA15510 and NA10851 [130].

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S3.CSV\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S3.CSV)

Additional file 4

Plots of the 79 CNV sites. Plots of the 79 sites of copy number difference between HapMap samples NA15510 and NA10851 (as listed in Additional File 3).

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S4.PDF\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S4.PDF)

Additional file 5

Pathological and clinical summaries for the 6 tumours. Details of the sample identity and cellularity composition as well as the construction of the pooled normal sample and the dilutions.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S5.CSV\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S5.CSV)

Additional file 6

Anticipated aberrations and known copy number changes in various samples. A list of known copy number changes for the cell-lines (SUM159, MT3, NA15510 and NA10851) and anticipated copy number changes for the tumours.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S6.PDF\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S6.PDF)

Additional file 7

Image plots of BASH processed Illumina data. False-colour image representation of six different raw images from the Illumina dataset that had significant spatial artefacts as identified using the BASH method from the beadarray Bioconductor package. As BeadStudio does not take spatial information into account during pre-processing, the resultant summarized values may be compromised in the presence of such artefacts.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S7.PNG\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S7.PNG)

Additional file 8

Experimental design for the Affymetrix platform. Targets file detailing the sample hybridized to each Affymetrix array.

[Click here for file](#)

[\[http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S8.XLS\]](http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S8.XLS)

Additional file 9

Experimental design for the Agilent platform. Targets file detailing the samples hybridized to each Agilent array.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S9.XLS>]

Additional file 10

Experimental design for the Illumina platform. Targets file detailing the sample hybridized to each Illumina array.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S10.XLS>]

Additional file 11

Experimental design for the Nimblegen platform. Targets file detailing the samples hybridized to each Nimblegen array.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S11.XLS>]

Additional file 12

All sample/chromosome plots for the tumours. Zip folder containing PNGs of all whole-chromosome plots for the tumours.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S12.ZIP>]

Additional file 13

All sample/chromosome plots for the cell-lines. Zip folder containing PNGs of all whole-chromosome plots for the cell-lines.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-588-S13.ZIP>]

Acknowledgements

We acknowledge the support of the University of Cambridge, Cancer Research UK, and Hutchinson Whampoa. ST is a Royal Society-Wolfson Research Merit Award holder. MJD was supported in part by a grant from the Medical Research Council. We thank Michelle Osborne and Sarah Moffatt for technical assistance with the Agilent and Illumina hybridizations. We also thank Andrew Teschendorff and Sergii Ivakhno for constructive discussions, Gulisa Turashvili for pathological expertise, and Matthew Hurles for access to CNV validation data.

References

- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR and Chen W, et al: **Global variation in copy number in the human genome.** *Nature* 2006, **444**(7118):444-454.
- Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, Lowry S, Gordon LA, Scott D, Xie G and Huang W, et al: **The DNA sequence and comparative analysis of human chromosome 5.** *Nature* 2004, **431**(7006):268-274.
- Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061-1068.
- Gaasenbeek M, Howarth K, Rowan AJ, Gorman PA, Jones A, Chaplin T, Liu Y, Bicknell D, Davison EJ and Fiegler H, et al: **Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers.** *Cancer Res* 2006, **66**(7):3471-3479.
- Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, Nathanson K, Protopopov A, Weber BL and Chin L: **A comparison of DNA copy number profiling platforms.** *Cancer Res* 2007, **67**(21):10173-10180.
- Coe BP, Ylstra B, Carvalho B, Meijer GA, Macaulay C and Lam WL: **Resolving the resolution of array CGH.** *Genomics* 2007, **89**(5):647-653.
- Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG and Veltman JA: **Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis.** *DNA Res* 2007, **14**(1):1-11.
- Gunnarsson R, Staaf J, Jansson M, Ottesen AM, Goransson H, Liljedahl U, Ralfkiaer U, Mansouri M, Buhl AM and Smedby KE, et al: **Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukemia—a comparative study of four differently designed, high resolution microarray platforms.** *Genes Chromosomes Cancer* 2008, **47**(8):697-711.
- Baumbusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, Bruhn L, Gunderson K, Naume B, Kristensen VN and Liestol K, et al: **Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors.** *BMC Genomics* 2008, **9**:379.
- Wicker N, Carles A, Mills IG, Wolf M, Veerakumarasivam A, Edgren H, Boileau F, Wasyluk B, Schalken JA and Neal DE, et al: **A new look towards BAC-based array CGH through a comprehensive comparison with oligo-based array CGH.** *BMC Genomics* 2007, **8**:84.
- Bengtsson H, Ray A, Spellman P and Speed TP: **A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods.** *Bioinformatics* 2009, **25**(7):861-867.
- Kloth JN, Oosting J, van Wezel T, Suzhai K, Knijnenburg J, Gorter A, Kenter GG, Fleuren GJ and Jordanova ES: **Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex genetic alterations in cervical cancer.** *BMC Genomics* 2007, **8**:53.
- Garnis C, Coe BP, Lam SL, MacAulay C and Lam WL: **High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples.** *Genomics* 2005, **85**(6):790-793.
- Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK and Kennedy GC, et al: **A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays.** *Cancer Res* 2005, **65**(14):6071-6079.
- Beroukhi R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK and Hofer MD, et al: **Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays.** *PLoS Comput Biol* 2006, **2**(5):e41.
- Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Hoglund M, Borg A and Ringner M: **Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios.** *BMC Bioinformatics* 2008, **9**:409.
- Lynch AG, Dunning MJ, Iddawela M, Barbosa-Morais NL and Ritchie ME: **Considerations for the processing and analysis of GoldenGate-based two-colour Illumina platforms.** *Stat Methods Med Res* 2009.
- Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG and Dermizakis ET, et al: **Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization.** *Genome Biol* 2007, **8**(10):R228.
- Scherer SW, Lee C, Birney E, Altschuler DM, Eichler EE, Carter NP, Hurles ME and Feuk L: **Challenges and standards in integrating surveys of structural variation.** *Nat Genet* 2007, **39**(7 Suppl):S7-S15.
- Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E and Chory J: **Large-scale identification of single-feature polymorphisms in complex genomes.** *Genome Res* 2003, **13**(3):513-523.
- Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays.** *Nat Genet* 2007, **39**(7 Suppl):S16-S21.
- Illumina: **Infinium II Assay Workflow Whitepaper, Illumina SNP Genotyping.** 2006.
- Chromosome X Titration Data Set.** http://www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx.

24. Bengtsson H, Irizarry R, Carvalho B and Speed TP: **Estimation and assessment of raw copy numbers at the single locus level.** *Bioinformatics* 2008, **24(6)**:759–767.
25. Cairns JM, Dunning MJ, Ritchie ME, Russell R and Lynch AG: **BASH: a tool for managing BeadArray spatial artefacts.** *Bioinformatics* 2008, **24(24)**:2921–2922.
26. Dunning MJ, Smith ML, Ritchie ME and Tavaré S: **beadarray: R classes and methods for Illumina bead-based data.** *Bioinformatics* 2007, **23(16)**:2183–2184.
27. Forozan F, Mahlamaki EH, Monni O, Chen Y, Veldman R, Jiang Y, Gooden GC, Ethier SP, Kallioniemi A and Kallioniemi OP: **Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data.** *Cancer Res* 2000, **60(16)**:4519–4525.
28. Davidson JM, Gorringe KL, Chin SF, Orsetti B, Besret C, Courtay-Cahen C, Roberts I, Theillet C, Caldas C and Edwards PA: **Molecular cytogenetic analysis of breast cancer cell lines.** *Br J Cancer* 2000, **83(10)**:1309–1317.
29. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P and Leal SM, et al: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449(7164)**:851–861.
30. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D and Pinkel D, et al: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37(7)**:727–732.
31. Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C, Clark R, Dovey O, Ellis P and Feuk L, et al: **Accurate and reliable high-throughput detection of copy number variation in the human genome.** *Genome Res* 2006, **16(12)**:1566–1574.
32. Conrad DF, Andrews TD, Carter NP, Hurles ME and Pritchard JK: **A high-resolution survey of deletion polymorphism in the human genome.** *Nat Genet* 2006, **38(1)**:75–81.
33. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C and Daly MJ, et al: **Common deletion polymorphisms in the human genome.** *Nat Genet* 2006, **38(1)**:86–92.
34. Hambly RJ, Double JA, Thompson MJ and Bibby MC: **Establishment and characterisation of new cell lines from human breast tumours initially established as tumour xenografts in NMRI nude mice.** *Breast Cancer Res Treat* 1997, **43(3)**:247–258.
35. **Epithelial Cancer Cell Lines database.** <http://www.path.cam.ac.uk/~pawefish/index.html>.
36. Maniatis T, Fritsch EF and Sambrook J: **Molecular cloning: a laboratory manual.** Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory; 1982.
37. Carvalho B, Bengtsson H, Speed TP and Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.** *Biostatistics* 2007, **8(2)**:485–499.
38. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A and Smyth GK: **A comparison of background correction methods for two-colour microarrays.** *Bioinformatics* 2007, **23(20)**:2700–2707.
39. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article 3.
40. R Development Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria; 2008.
41. Edgar R, Domrachev M and Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30(1)**:207–210.
42. Durinck S, Bullard J, Spellman PT and Dudoit S: **GenomeGraphs: integrated genomic data visualization with R.** *BMC Bioinformatics* 2009, **10**:2.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression

ANNA GIT,^{1,3} HEIDI DVINGE,^{2,3} MALI SALMON-DIVON,^{2,3} MICHELLE OSBORNE,¹ CLAUDIA KUTTER,¹ JAMES HADFIELD,¹ PAUL BERTONE,² and CARLOS CALDAS¹

¹Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom

²European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, United Kingdom

ABSTRACT

RNA abundance and DNA copy number are routinely measured in high-throughput using microarray and next-generation sequencing (NGS) technologies, and the attributes of different platforms have been extensively analyzed. Recently, the application of both microarrays and NGS has expanded to include microRNAs (miRNAs), but the relative performance of these methods has not been rigorously characterized. We analyzed three biological samples across six miRNA microarray platforms and compared their hybridization performance. We examined the utility of these platforms, as well as NGS, for the detection of differentially expressed miRNAs. We then validated the results for 89 miRNAs by real-time RT-PCR and challenged the use of this assay as a “gold standard.” Finally, we implemented a novel method to evaluate false-positive and false-negative rates for all methods in the absence of a reference method.

Keywords: miRNA; microRNA; differential expression; microarray; real-time PCR; sequencing; pyrosequencing; miRNA-seq

INTRODUCTION

MicroRNAs (miRNAs) are regulatory noncoding RNA molecules ~20–23 nucleotides (nt) long, generated by two cleavage events mainly from RNA Pol II primary transcripts (pri-miRNAs) via a ~70-nt imperfect stem-loop intermediate (pre-miRNA). Over 10,000 miRNAs from 115 species, ranging from vertebrates (Lagos-Quintana et al. 2001) to viruses (Pfeffer et al. 2004), are currently deposited in the miRNA registry (miRBase version 14) (Griffiths-Jones et al. 2008). These include ~700 (out of up to ~3400 predicted) (Sheng et al. 2007) human miRNAs.

miRNAs mediate the translational repression, and sometimes degradation, of target mRNAs mostly by directing an RNA-induced silencing protein complex to imperfect complementary sequences in their 3'UTRs (van den Berg et al. 2008). Up to ~60% of human genes are putative targets of

one or more miRNA (Friedman et al. 2009). miRNAs play a role in all major biomolecular processes, including metabolism (Krutzfeldt and Stoffel 2006), cell proliferation (Bueno et al. 2008) and apoptosis (Jovanovic and Hengartner 2006), development and morphogenesis (Stefani and Slack 2008; He et al. 2009), stem cell maintenance, and tissue differentiation (Shi et al. 2006). miRNAs are reported to be involved in 94 human diseases (Jiang et al. 2009), ranging from psychiatric disorders (Barbato et al. 2008) through diabetes (Hennessy and O'Driscoll 2008) to cancer (Medina and Slack 2008).

Three principal methods are used to measure the expression levels of miRNAs: real-time reverse transcription-PCR (qPCR) (Chen et al. 2005; Shi and Chiang 2005), microarray hybridization (Yin et al. 2008; Li and Ruan 2009), and massively parallel/next-generation sequencing (NGS) (Hafner et al. 2008), all of which face unique challenges compared to their use in mRNA profiling. In terms of microarray analysis, the short length of mature miRNA sequences constrains probe design, such that often the entire miRNA sequence must be used as a probe. Consequently, the melting temperatures of miRNA probes may vary by >20°C. qPCR assays, traditionally relying on the

³These authors contributed equally to this work.

Reprint requests to: Anna Git, Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom; e-mail: Anna.Git@cancer.org.uk; fax: 44-1223-404199.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.1947110>.

specificity provided by a number of contiguous probes, compensate for the compromised sequence specificity by a stringent spatial constraint (3' terminal or near-terminal sequences). A similar constraint is also imposed by stem-loop microarray probes (Agilent). NGS of miRNAs can be influenced by sequencing errors and often requires search and removal of adaptor sequences before the miRNA sequence itself can be elucidated.

A second challenge in measuring miRNA levels arises from the existence of miRNA families, the largest encompassing nine variants (*hsa-let-7a-i*), whose members differ by as little as one nucleotide but nevertheless exhibit differential expression patterns (Roush and Slack 2008). The stringency required to differentiate between these closely related miRNA species surpasses that of conventional mRNA microarrays. This challenge is partly addressed by ensuring that hybridization-based assays are performed at high enough temperatures to reject cross-hybridizing transcripts. In addition, microarrays with probes containing locked nucleic acid (LNA) bases (Exiqon) provide higher annealing affinities, potentially allowing the assay to discern between individual miRNA family members and somewhat equalizing the melting temperatures of probe sequences. Finally, advances in sequencing technology have accelerated both the discovery rate of new miRNAs and modifications to existing miRNA entries, reflecting subtle variations in mature miRNA sequences (e.g., post-transcriptional editing or terminal residue addition) (Landgraf et al. 2007). As a result, the continued refinement of miRNA databases necessitates frequent changes to miRNA array probe design and annotation.

The technical merits and drawbacks of qPCR, microarrays, and sequencing of miRNAs are similar to their application for RNA or genomic DNA quantitation. The clear advantage of high-throughput sequencing is the ability to identify novel miRNAs. This technology is not hindered by variability in melting temperatures, coexpression of nearly identical miRNA family members, or post-transcriptional modifications. However, both the RNA ligation (Bissels et al. 2009) and the PCR amplification (see below) steps bear inherent biases, the method is laborious and costly, and associated tools for computational analysis are in their infancy. qPCR is often considered a “gold standard” in the detection and quantitation of gene expression. However, the rapid increase in number of miRNAs renders qPCR inefficient on a genomic scale, and it is probably better used as a validation rather than a discovery tool.

As with genomic DNA and RNA analysis, microarrays are still the best choice for a standardized genome-wide assay that is amenable to high-throughput applications. Over 400 existing publications have utilized commercial or in-house printed miRNA microarrays. The differences between available platforms range from surface chemistry and printing technology, through probe design and labeling techniques, to cost. Unlike for mRNA gene expression (Shi et al. 2006), comparative genomic hybridization (Baumbusch et al. 2008),

or chromatin immunoprecipitation (Johnson et al. 2008) assays, few attempts have been made to establish rigorous parameters for the evaluation of a miRNA microarray platform, especially in light of the specific challenges miRNAs present.

We have undertaken a systematic comparison of six commercially available miRNA microarray platforms representing single- and dual-channel fluorescence technologies, using three well-defined RNA samples (Git et al. 2008), and compared the results with NGS and qPCR. This study represents, to the best of our knowledge, the most comprehensive comparison of the performance of methods to detect differentially expressed miRNAs to date.

RESULTS

Microarray comparison study design

As a preface to this study, we extensively evaluated RNA extraction and quality control (QC) methods to ensure a high standard of quality for the RNA samples used (data not shown). The biological samples were representative of a realistic application of miRNA microarrays in a cancer research institute. Moreover, biological replicates of these samples have been previously profiled by contact-printed and bead-based microarrays (Git et al. 2008), providing a comparative reference for QC during preliminary stages as well as in final analyses (e.g., tumor suppressor [TS] miRNAs) (see Fig. 3B, below).

Three samples were analyzed in this study: a pool of commercial RNAs from normal breast tissue (N), the luminal breast cancer cell line MCF7 (M), and a breast progenitor cancer cell line PMC42 (P), which exhibits many normal-like characteristics. All three samples, extracted in bulk and quality assured, were labeled and hybridized in quadruplicate to six commercially available microarray platforms in strict accordance with the protocols recommended by the manufacturers (see Materials and Methods). The microarray platforms used in this study were the Agilent Human miRNA Microarray 1.0; Exiqon miRCURY LNA microRNA Array, v9.2; Illumina Sentrix Array Matrix 96-well MicroRNA Expression Profiling Assay v1; Ambion mirVana miRNA Bioarrays v2; Combimatrix microRNA 4X2K Microarrays; and Invitrogen NCode Multi-Species miRNA Microarray v2. For simplicity, each platform is referenced throughout after its manufacturer name.

Microarray hybridization performance

Figure 1 depicts the distributions of several measures of hybridization quality and consistency, such as the signal-to-noise ratio (SNR) (Fig. 1A), the coefficient of variation (CV) between replicate spots and arrays (Fig. 1B,C), and pairwise correlation between arrays (Fig. 1D). Overall, the SNR generated by the Normal samples was the highest, and MCF7 was the lowest for each platform evaluated. This

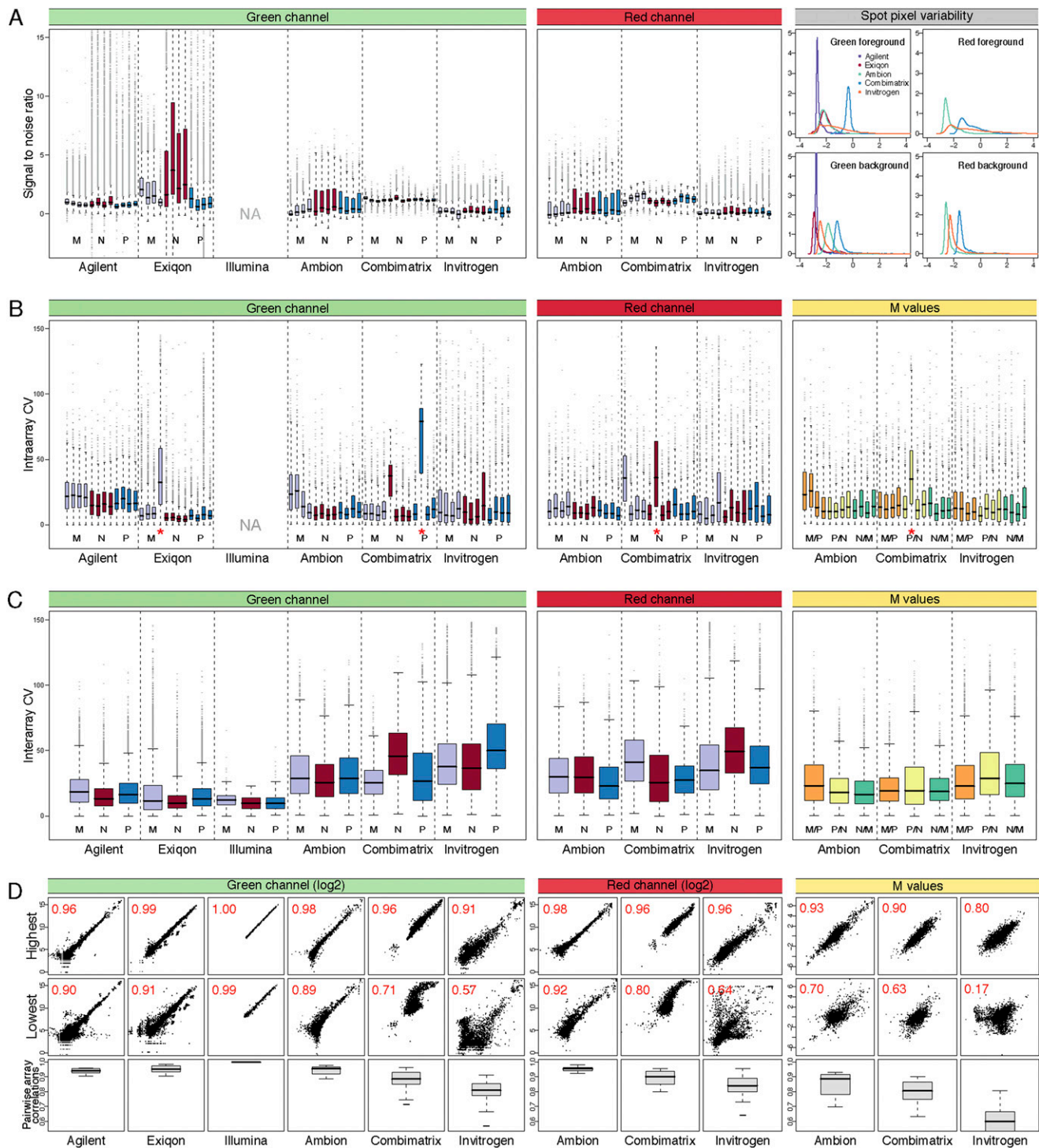


FIGURE 1. Analysis of hybridization performance. (A) Signal-to-noise ratio for the raw 532 nm/Cy3 (green banner) and 635 nm/Cy5 (red banner) intensities for all spots on the individual arrays was calculated using the SSDR method. For Illumina arrays, this calculation was impossible as only the foreground intensities were available. Purple indicates arrays with M samples; red, N, and blue, P. For clarity of presentation, the y-axis was truncated at 15, thereby excluding some extreme outliers. The distribution of the log₂ standard deviation between pixels within each spot scaled to the median spot intensity is shown on the *right* (gray banner). (B) Intra-array coefficients of variation across replicated spots on each array were calculated for the unprocessed Cy3 and Cy5 intensities (bar and banner colors as above), and the log₂ ratios (M-values, yellow banner; orange bars indicates M/P; yellow; P/N, green, N/M). Arrays with a red asterisk were excluded from subsequent analysis. (C) Interarray coefficients of variation were calculated for arrays hybridized with the same samples (bar and banner colors as above). (D) Pairwise correlations for arrays hybridized with the same samples were calculated (15–18 correlations). Distribution of R² values are shown in box plots (*bottom* row), with the highest (*top* row) and lowest (*middle* row) correlations shown as examples. The axis for the *bottom* row was truncated at 0.55 for clarity, excluding some of the values for Invitrogen.

agrees with the observation that overall miRNA content is reduced in cell lines compared to tissue (Lee et al. 2008; C Blenkiron and LD Goldstein, pers. comm.). PMC42, a normal-like cell line, demonstrated intermediate levels of overall miRNA expression. The difference was least pronounced on platforms with a high within-spot pixel variability (Combimatrix, Invitrogen) (Fig. 1A, right panel), since the SNR not only depends on fluorescent signal intensities but is inversely proportional to the standard deviation of both foreground and background pixels, so that high spot uniformity contributes to higher SNR. Some typical spot artifacts leading to low uniformity were indeed observed during feature extraction (“doughnut-shaped” spots for Invitrogen indicate high signal on the outside of spots, low on the inside, and the opposite pattern for Combimatrix) (data not shown).

We then examined the variability between replicates spotted on the same array (intra-array CV) (Fig. 1B). For dual-channel arrays, the 532-nm (Cy3, green) and 635-nm (Cy5, red) fluorescence intensities and their \log_2 ratios (M-values) were treated separately, since localized signal variations occurring in both channels may cancel each other out. We observed no consistent differences between single- and dual-color platforms, and although CV's varied considerably between some platforms, they tended to be consistent within platforms, with the exception of two arrays subsequently excluded from downstream analysis.

The interarray CV's (Fig. 1C) were calculated for each type of probe (several different probe types might target the same miRNA) across all replicate spots on the four replicate arrays. These typically include 12–24 values, although some probes (e.g., controls or empty spots) were present in greater numbers; for example, Agilent arrays contain more than 3000 empty spots. Single-channel hybridization was more consistent across replicates, as evident from the overall lower CV values, but these differences were ameliorated when the M-values, rather than the individual Cy3 and Cy5 intensities, were considered for the dual-color platforms. Reproducibility between hybridizations was also assessed by pairwise comparison of replicate arrays. The distribution of the resulting R^2 values as well as the most and least consistent examples from each platform, are illustrated in Figure 1D. Although all platforms demonstrated at least one replicate pair with greater than 0.9 (and usually >0.95) R^2 correlation, their distribution was much wider. Notably, unlike the interarray CV values, the dual-channel replicates with low correlation (below an R^2 of 0.8) showed poorer agreement when treated as M-values instead of Cy3 and Cy5 intensities. This may be due to the inaccuracy of M-values for low-intensity spots. In particular, negative control or empty replicate spots were considered individually for the pairwise comparison, thus strongly affecting the distribution of M-value correlations, but were condensed into single values across all interarray replicates for the interarray CVs (Fig. 1C).

We then proceeded to analyze the consistency of the detected spots for each platform. First, the frequency of

“present/marginally present” or “absent” calls was calculated for each spot on the arrays based on the intensity of negative controls and empty spots (see Materials and Methods) (Fig. 2A). The platforms varied significantly in the consistency of associated present/absent calls, visually represented by the thickness of the “belt” region separating the red (consistently “present”) and blue (consistently “absent”) zones of the bars: whereas the “belt” values comprised fewer than 20% of the probes in Agilent, they accounted for over half the probes in Invitrogen arrays. This variation stems from interarray variability and the availability of spots to evaluate the background distribution. For example, despite the very similar interarray CV of the M-values in Ambion and Combimatrix assays (Fig. 1C), the consistency of calls on the Ambion array platform was higher.

Microarray probe mapping and hybridization specificity

Due to the inherent difficulties associated with miRNA probe design outlined in the introduction, the complements of miRNAs targeted by each platform are difficult to compare. To allow an accurate comparison between the platforms, we reannotated all the probes against miRBase version 12 using uniform criteria (see Materials and Methods). Although the total number of probes varied significantly across platforms, the number of human miRNAs represented on the array was fairly constant and depended mainly on the miRBase version at the time of array design. The overall characteristics of probes represented in each platform and the effect of re-annotation are summarized in the “probe properties” section of Table 1.

Reannotated probes were divided into categories based on information from the manufacturers and our remapping of probe sequences (see Materials and Methods). The categories are listed and color coded in the legend to Figure 2, B and C, according to our approximate expectation regarding their intensity (high/red to low/blue). Of particular interest were human miRNAs and potential cross-hybridizing probes (mouse miRNAs and probes with mismatches to human miRNAs; MM_human). We counted the number of probes called as “present” in each category (Fig. 2B) and examined the distribution of their normalized signal intensities (Fig. 2C). Platforms varied both in overall signal intensity and number of probes called “present.” The former property is affected by a combination of labeling chemistry, input RNA concentration, and hybridization efficiency, with Combimatrix arrays producing the brightest signal. However, the low numbers of “present” calls (127, 85, 105) on this platform are similar to those produced by the low-intensity Invitrogen arrays (49, 103, 100), underscoring the importance of distinguishing between the two metrics.

Within most platforms the signal and “present” rate also varied extensively depending on spot category. As expected, positive controls were usually among the brightest spots, and

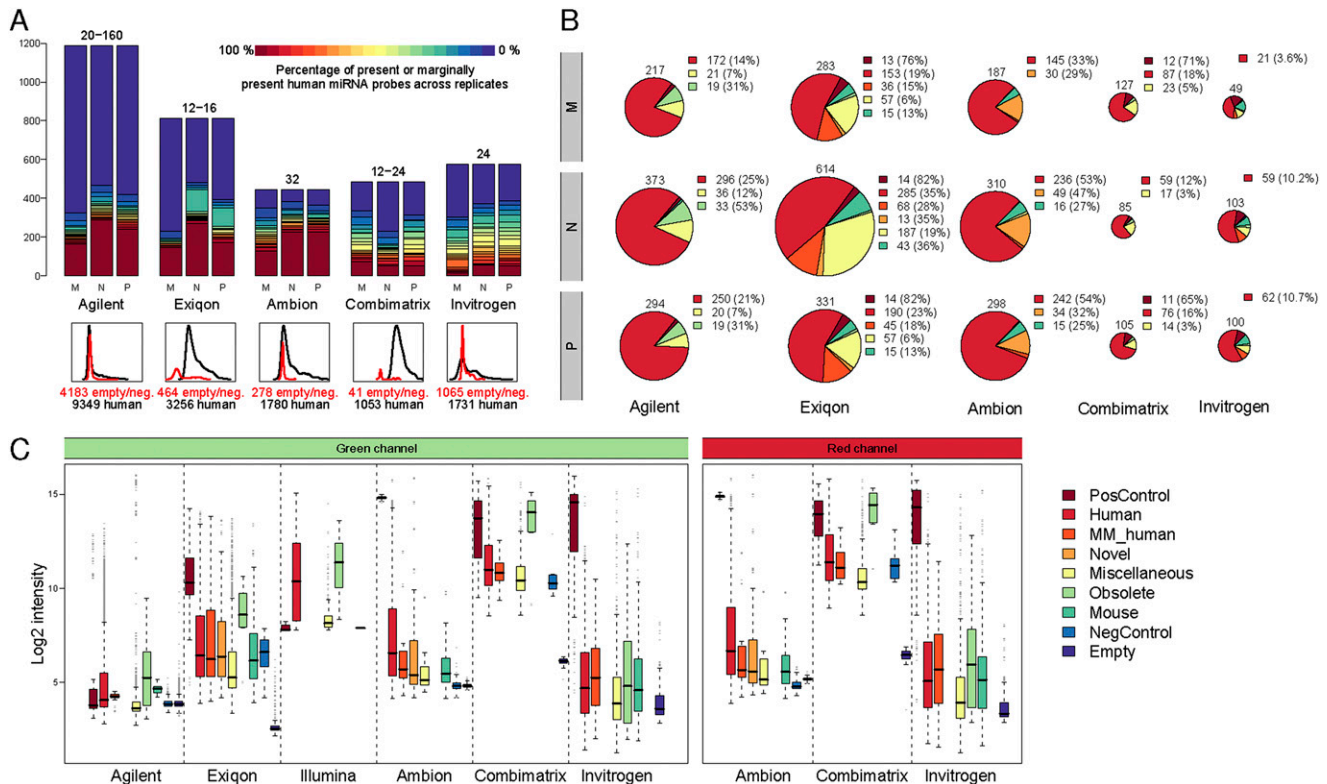


FIGURE 2. Analysis of detected probes. (A) Consistency of present/absent calls among human miRNAs. (Top) For each human probe, the percentage of replicates detected (called present) by the platform was calculated and summarized (bars). The numbers above the bars indicate number of probe replicates. (Bottom) Intensity distribution of human miRNAs (black) and the empty and negative spots used to calculate the nonspecific binding (red), with the number of probes of each type listed below the plot. Illumina array data are missing from panels A and B, as information regarding negative or empty spots was not available. (B) Detected spot types. Probes have been categorized based on their target miRNAs (see Materials and Methods). The number of unique spots from each category being detected as “present” in >90% of its replicates across all arrays was calculated for each of the three samples types. For categories with 10 or more present probes, the count is shown next to the figure, with the proportion of the “present” calls out of the total probes in that category (%). The radius of each chart is proportional to the total number of present spots, indicated above. The legend is shared with panel C. PosControl and NegControl are positive and negative controls, respectively; MM_human, mismatched human. (C) Intensity range of the different spot types. For each of the spot types of panel B, the distribution of intensities of background-corrected and normalized green or red log₂ values across all arrays was calculated.

probes targeting human miRNAs had the broadest range representing varying levels of expression or tissue specificity of miRNAs. Probes matching mouse miRNAs or MM_human miRNAs were clearly “present” in some platforms and not others, indicating a degree of cross-hybridization between similar probes. For example, the intensities of mouse probes or mismatched probes in Agilent did not differ greatly from the negative or empty probes, and indeed, less than 10 probes were called “present” in each category. In contrast, the spread of Exiqon intensities in the mouse and human_MM categories was large, and in the Normal sample, “present” calls were made for 68 mismatched probes and 43 mouse probes, representing 28% and 36% of the total number of probes in the respective category. We note that most of the mismatched probes identified by our uniform reannotation are classified by Exiqon as “obsolete” or “not_designed_for_hsa,” so they may not exhibit the same LNA spiking pattern as their perfect-match counterparts. The distribution of signals upon removal of these probes is

available in Supplemental file “Exiqon annotation.” Probe specificity is evaluated by Exiqon using synthetic RNA spike-ins in a relatively low complexity background (yeast tRNA). The biological relevance of the two analyses (mismatched probes versus spike-ins) remains to be elucidated.

Worthy of mention is the considerable number of “present” calls made by most platforms in other categories, such as miscellaneous or obsolete, and in particular the novel category in Ambion containing proprietary “Ambi-miRs.” These results emphasize the limited information offered by overall signal intensities or total number of detected features, often quoted as measures of hybridization performance and platform sensitivity.

Correlation of microarray and NGS data

We proceeded to sequence the mature miRNAs from each of the samples using a Genome Analyzer II platform (Illumina; hereafter abbreviated as GAsq). On average,

TABLE 1. Comparison of the practical aspects of RNA handling and hybridization and probe properties

	Agilent	Exiqon	Illumina	Ambion	Combinatrix	Invitrogen
General Array version used	miRNA Microarray 1.0	miRCURY LNA microRNA Array, v9.2	Sentrix Array Matrix 96-well MicroRNA Expression Profiling Assay v1	mirVana miRNA Bioarrays v2	microRNA 4X2K Microarrays	NCode Multi-Species miRNA Microarray v2
Channels	Single	Dual (used here as single due to low Cy5 intensities)	Single	Dual	Dual	Dual
RNA input amount (ng)	100	1000	200	1000	1000	1000
RNA input type	Total	Total	Total	sRef	sRef	Total
Labeling chemistry	CIP and 3' ligation	CIP and 3' ligation	Polyadenylation and RT-annealing-PCR	Polyadenylation and coupling	Direct coupling of dye to G residues	Limiting polyadenylation and 3' splint ligation
Arrays per slide	8	1	96-well	2	4	1
Ease of handling	++; column cleanup;	+++; freezing not recommended	+: multi-day protocol	+: multi-day protocol with several drying and cleanup steps	++; column cleanup; freezing not recommended	++
Hybridization and washes	+++; drying before hybridization	+++	+	+: drying before hybridization;	+++	++; drying before hybridization;
Feature extraction (printing uniformity)	+++	++	NA	+	+++	++
Probe properties						
Probe type	Stem-loop with active sequence	Contains LNA-modified bases		miRNA sequence with spacer	miRNA sequence	Tandem repeat
Total unique probes (spots per array)	1589 (15744)	2245 (9792)	743	662 (2856)	1032 (2240)	1146 (4608)
Empty probes (spots per array)	1 (3511)	1 (344)	—	1 (194)	1 (34)	290 (1065)
Human probes (spots per array)	1187 (9349)	814 (3256)	485	445 (1780)	483 (1053)	577 (1731)
Mismatched human probes (spots per array)	3 (25)	159 (636)	—	17 (68)	10 (22)	62 (186)
Miscellaneous probes (spots per array)	308 (1575)	1066 (4384)	244	14 (56)	509 (1095)	344 (1053)
Mouse probes (spots per array)	3 (30)	135 (540)	—	59 (236)	—	133 (399)
Negative control probes (spots per array)	7 (672)	10 (120)	2	21 (84)	7 (7)	—
Novel probes (spots per array)	—	38 (152)	—	105 (420)	—	—

(continued)

TABLE 1. Continued

	Agilent	Exiqon	Illumina	Ambion	Combinatrix	Invitrogen
Obsolete probes (spots per array)	62 (401)	6 (24)	8	0	6 (12)	14 (42)
Positive control probes (spots per array)	19 (181)	17 (336)	4	1 (18)	17 (17)	16 (132)
Spotted replicates (number of probes)	5 (805), 6 (85), 7 (168), <u>10 (503)</u> , 26 (1), 34 (1), 40 (19), 64 (6), 288 (1), 3511 (1)	4 (2233), 28 (1), 32 (9), 40 (1), 160 (1), 344 (1)	NA	4 (661), 18 (1), 194 (1)	1 (32), 2 (826), <u>3 (174)</u> , 34 (1)	3 (1364), 6 (70), 24 (1), 72 (1)
Distribution of replicates	Irregular	Regular		Regular	Irregular	Regular
Probe length (number of probes)	9 (3), 10 (6), 11 (17), 12 (41), 13 (77), 14 (119), 15 (173), 16 (214), <u>17 (233)</u> , 18 (219), 19 (192), 20 (123), 21 (57), 22 (26), 23 (8), 24 (1), 25 (1)	10 (4), 11 (10), 12 (21), 13 (56), 14 (93), 15 (145), 16 (168), 17 (198), 18 (266), 19 (241), 20 (243), <u>21 (417)</u> , 22 (269), 23 (79), 24 (29), 25 (2), 26 (3)	16 (13), 17 (177), <u>18 (249)</u> , 19 (127), 20 (85), 21 (66), 22 (15), 23 (3)	16 (1), 17 (7), 18 (6), 19 (8), 20 (34), 21 (129), <u>22 (315)</u> , 23 (116), 24 (22), 25 (4), 26 (1), 27 (19)	17 (4), 18 (6), 19 (40), 20 (52), 21 (249), <u>22 (406)</u> , 23 (200), 24 (47), 25 (25), 26 (2), 28 (2), 33 (1), 34 (1), 35 (38), 40 (1)	24 (1), 30 (1), 32 (3), 34 (9), 36 (21), 38 (51), 40 (128), 42 (387), <u>44 (540)</u>
Coverage of human miRNAs (listed/reannotated)	470/450	572/599	470/480	328/371	470/463	467/483
miRBase design version	9.2?	9.2		8		9
Probes with several targets (potential cross hybridization)	25	22	14	25	22	25
Average number of detected probes	295	409	NA	265	106	84

Where multiple values are listed within a cell, the most frequent ones are underlined; NA, nonapplicable.

12 million reads were obtained in each sequencing run, and after filtering, 32%, 35%, or 63% (M, P, and N, respectively) could be mapped to known miRNAs. Overall, 733 miRNAs were detectable (501 in M, 588 in P, and 608 in N), and 472 of those had at least 10 cumulative counts across the three samples. The number of reads obtained for each miRNA was well-correlated to the respective microarray hybridization intensity (Pearson correlation 0.66 ± 0.12 , ranging between 0.42 and 0.87; see Supplemental file “Intensity Correlations”). The 45 miRNAs that were not

identified in the sequencing data set, but for which expression levels were in the detection range of at least one microarray platform, were typically called “marginally present” in the latter, suggesting low cross-hybridization of the corresponding array probes.

To allow a direct comparison of the platforms’ performances, we focused on the intersection of miRNAs represented on all platforms (Fig. 3A). A total of 215 miRNA probes were included on only one microarray platform (most often Exiqon), while 148 miRNAs were represented

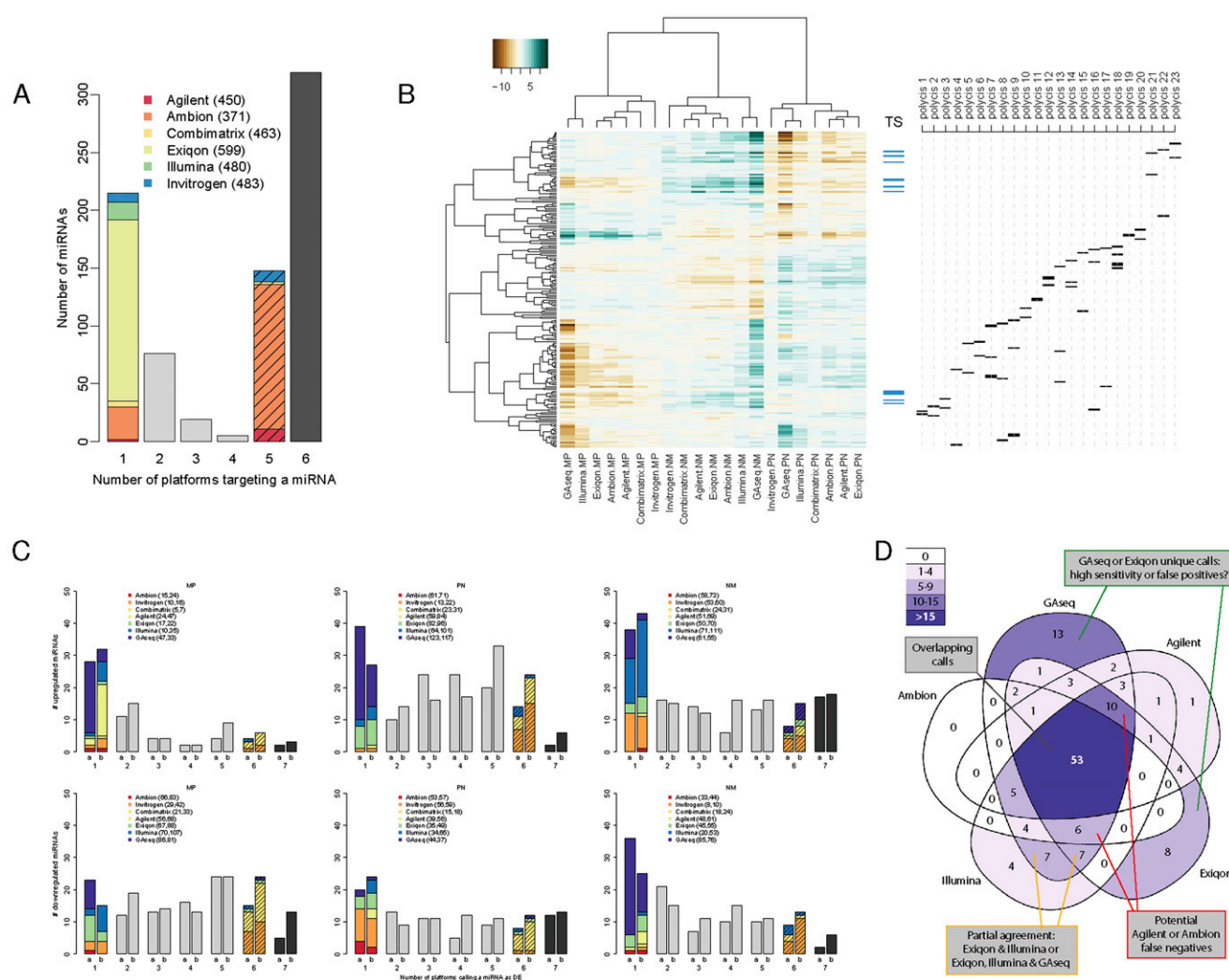


FIGURE 3. Analysis of differential expression. (A) miRNA targeting by platforms. The number of reannotated miRNAs targeted by varying numbers of platforms was calculated. Solid colors indicate miRNAs found only on the indicated platform; striped colors, miRNAs found on all platforms *except* the indicated platform. The total number of human miRNAs on each platform is indicated in parenthesis. Black bar indicates 319 miRNAs represented on all microarrays. (B) Clustering of the common probe M-values. M-values of 204 human probes common to all microarray platforms with no predicted cross-hybridization and detectable by Gaseq were subjected to unsupervised clustering using Pearson correlation. Ticks indicate the position of potential tumor suppressor (TS) miRNAs (blue) and miRNAs arising from a single genomic location contained in a putative polycistronic pri-miRNA (black). A list of polycistrons is provided in Supplemental file “Polycistrons.” (C) Consistency of DE calls by all platforms. The number of platforms calling each miRNA as DE (up-regulated, *top*; down-regulated, *bottom*) in each of the three biological comparisons was recorded. DE calls were derived (1) using a uniform threshold of \log_2 fold-change >1 or (2) using optimal thresholds calculated for each platform by the iMLE algorithm. The overall number of relevant DE calls made by each platform is indicated in parenthesis. (D) Overlap in DE calls of five platforms. The number of miRNAs called by five platforms as up-regulated in P versus N sample using iMLE-optimized cutoffs was plotted inside a Venn diagram. Areas are shaded according to number of DE calls and their relative sizes bear no meaning.

on five out of the six microarrays. This was predominantly due to their absence from the Ambion arrays, which were designed against an earlier version of miRBase. Three hundred nineteen miRNAs were targeted by all six microarray platforms; of these, 204 had no predicted cross-hybridization and at least 10 GA sequencing reads mapped to mature sequences.

For these 204 miRNAs, the \log_2 ratios were calculated for the M/P, P/N, and N/M comparisons and clustered based on Pearson correlation (Fig. 3B). Importantly, as this analysis is limited to a subset of miRNAs, it should not be considered as a complete comparison of the three biological samples. All M-values clustered according to the biological comparison rather than platform type. Data obtained from the two PCR-based methods (GAseq sequencing and Illumina microarrays) consistently clustered together, as did the data from the three microarray platforms exhibiting greater reproducibility (Exiqon, Ambion, and Agilent). The clustering of Invitrogen and Combimatrix data was inconsistent.

Recent reports have demonstrated the effect of normalization on the interpretation of miRNA expression data (Hua et al. 2008; Pradervand et al. 2009), which is certainly magnified by combining data from several platforms. We therefore tested whether our clustered normalized data reflected the coregulation of biologically meaningful groups, in particular potential TS miRNAs frequently lost in cancer (Git et al. 2008, and references therein), and groups of miRNAs residing in close genomic proximity and potentially cotranscribed as a polycistronic pri-miRNA. Figure 3B (right) indicates the relative positions of these miRNAs in the overall clustering, clearly demonstrating correlated levels of both potential TS miRNAs and the miRNA products of many putative polycistronic transcripts. Among the latter category, those groups that do not demonstrate coregulation may not in fact be polycistronic or may be individually regulated by post-transcriptional mechanisms.

We identified the differentially expressed genes on each platform using a uniform arbitrary fold-change threshold of 2 and corrected *P*-values of <0.05 (Fig. 3C, bars coded “a”) and examined the agreement between platforms. Surprisingly, the actual overlap between the differentially expressed (DE) calls of the platforms was very low. Consistent with the low rate of “present” calls, Invitrogen and Combimatrix results were most frequently in disaccord with the other microarray platforms, while GAsseq, Illumina, and Exiqon assays produced the highest numbers of unsupported DE calls.

To eliminate the possibility that the low degree of overlap between platforms resulted from applying an arbitrary uniform cutoff, we developed a novel iterative maximal likelihood estimate (iMLE) algorithm to establish the optimal cutoff for each platform in view of the combined data of all platforms. The overlap of the resulting DE calls is presented in bars “b” in Figure 3C. Although the optimized cutoffs increased the number of the fully overlapping DE

calls in all six sample comparisons, the vast majority of the DE calls were still not unanimous across platforms. Whether this disagreement ensues from nonspecific contributions, varying degrees of cross-hybridization of miRNA family members or reduced discrimination between unprocessed and mature forms of the miRNAs (only Agilent’s probes are mature-specific) is at present unknown and will necessitate the use of specific synthetic spike-in oligonucleotides.

The difference in DE calls for each comparison is the net result of sensitivity and specificity characteristics inherent to each platform, and those exhibiting the highest sensitivity are expected to make some unsupported DE calls and to generate increasingly large overlaps with platforms of lower sensitivity, evident as DE calls made by two to four platforms (Fig. 3C, gray bars). We therefore examined the nature of the overlap in DE calls. Figure 3D shows an example of the overlap between GAsseq and four of the six microarray platforms tested (Invitrogen and Combimatrix were excluded for ease of plotting) in identifying miRNAs up-regulated in the P/N comparison. Here, 53 miRNAs were called up-regulated by all platforms, and both GAsseq and Exiqon yielded a large number of unique DE calls (13 and eight, respectively), suggesting that at least one of the platforms exhibits high false-positive (FP) calls (i.e., reduced specificity). Similarly, 10 and six miRNAs were called significantly up-regulated by all platforms except Ambion and Agilent, respectively (false negatives [FNs]), indicative of lower sensitivity. In more complex overlap patterns, the same number (seven) of Exiqon and Illumina’s overlapping DE calls was supported or rejected by GAsseq. Since all platforms were given equal status, such data could not easily be translated into specificity (true negative [TN]) and sensitivity (true positive [TP]) values.

Correlation with qPCR results

Microarray and NGS data are regularly validated by qPCR. We analyzed the expression of 89 miRNAs from multiple overlap categories using either TaqMan or SYBR Green assays. The \log_2 ratios of this miRNA subset for all platforms were sorted according to the corresponding qPCR values (Fig. 4A). Although the trend of M-values follows that of the qPCR data, the magnitude of the M-value is clearly different between platforms (ratio compression). Occasional spurious values in single platforms are noticeable as red or blue “islands.”

The ratio compression can also be visualized by the slope of the concordance between each platform and qPCR data for each of the three biological comparisons, exemplified for GAsseq data in Figure 4B (average slope ~ 1 ; i.e., no compression). The average slopes for the microarray platforms are listed in Figure 4B and range between 0.24 (Invitrogen) and 0.61 (Ambion). Also evident in this plot is the shift in the *y*-axis intercept, representing a consistent drift in the measured ratio, also evident in microarray/qPCR plots

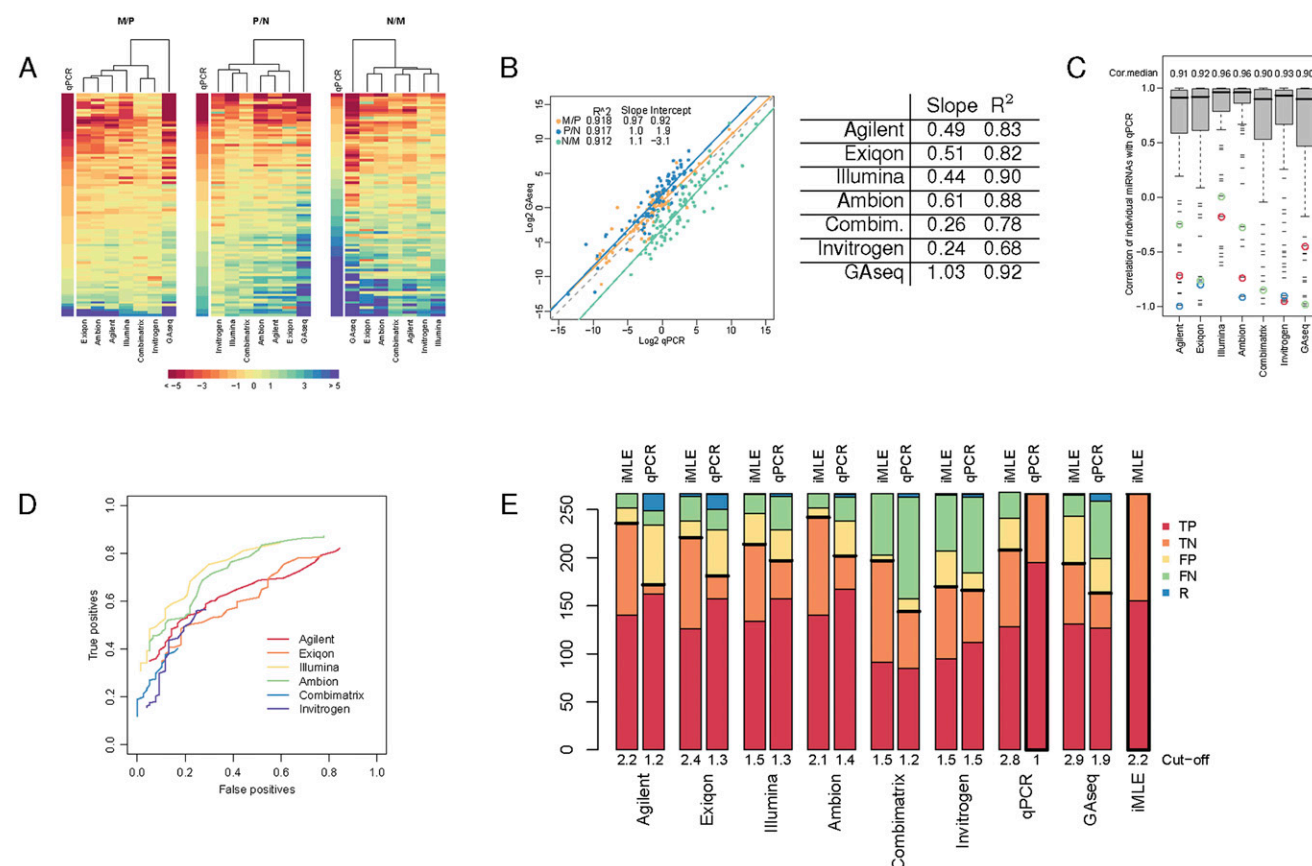


FIGURE 4. Validation by real-time RT-PCR. (A) M-values of miRNAs tested by qPCR. Eighty-nine miRNAs validated by qPCR (rows) are sorted by their qPCR M-values. Platforms (columns) are clustered by Euclidean distance. (B) Overall correlation between Gaseq and qPCR data. For each biological comparison, the ratios of miRNA expression calculated from Gaseq were plotted against those derived from qPCR. Best linear regression fit (solid lines; R² values, intercept with y-axis and slope indicated in legend); Y = X (dotted line). Average correlations and slopes across the three comparisons are listed for each platform compared to qPCR. (C) Correlation between microarray/NGS and qPCR data. Boxes depict the distribution of correlation for the M-values generated by qPCR and indicated platforms for each miRNA in all three comparisons (MP, PN, NM), and the median value (Cor.median) is indicated above. Examples of consistent outliers are circled; hsa-miR-484 (red), hsa-miR-15a (green), and hsa-miR-215 (blue). (D) Effect of DE cutoff on the TP and FP rate of each platform. The number of TP and FP DE calls, compared with qPCR calls at fold-change >2 was calculated across a range of thresholds (0–5 in 0.1 increments). Only miRNAs with P-value <0.05 were included for each platform; hence, the ROC curves do not cover the entire range of TP and FP rates. (E) True and false call rates of each platform at optimal cutoffs. The number of TP and FP and FN DE calls was calculated at the optimal log₂ cutoffs calculated based on a qPCR reference or on the iMLE algorithm with qPCR as an unknown platform. The number of DE (equivalent to TP) and non-DE (equivalent to TN) calls made by these references is shown with a thick frame. A horizontal black thick line separates true calls (below) from false calls (above). Abbreviations as in panel C.

(data not shown). This shift arises from the inherently different ratio of the overall miRNA population and external reference genes used in qPCR normalization (e.g., 5S rRNA). It has been repeatedly observed by ourselves and others (Lee et al. 2008; C Blenkiron and LD Goldstein, pers. comm.) that cell lines have a lower miRNA content per total RNA (>85% of which is rRNA species) than tissue samples. This trend is supported by the fact that despite similar quality and quantity of RNA input the overall hybridization signal in MCF7 arrays is lower than in normal breast tissue arrays (Fig. 1A) with the normal-like cell line PMC42 showing intermediate values. NGS and microarrays are for the most part blind to such fluctuations as they employ normalization techniques within the miRNA population. As a result, every miRNA appears

to be better-expressed in M samples when measured by Gaseq compared with equivalent qPCR measurements, where its levels are normalized to high 5S content, thus consistently shifting the Gaseq N/M ratios down (intercept = −3.1). Similarly, M/P and P/N correlations are shifted by +0.92 and +1.9, respectively.

The concordance of each platform with qPCR data was measured as either Pearson correlation of all array M-values against the matching qPCR M-values (comparing columns in a traditional table layout) (Fig. 4B, R² values) or the distribution of correlations of the M-values of individual miRNAs in the three comparisons (comparing rows in a traditional table layout) (Fig. 4C, box plots). The two measures do not necessarily agree (e.g., Invitrogen's median correlation is 0.93, although the overall average

correlation is only 0.68). Discrepancies could arise due to a relatively small number of poorly correlating outliers (counted once for box plots but strongly skewing an overall linear fit) or as a result of differences in the correlation slopes of individual probes (i.e., rows), which—while possibly scoring well in a box plot analysis—reduce the quality of the overall (i.e., columns) linear fit.

We then extended our analysis from continuous M-values to discrete DE calls. Using the calls generated by qPCR as a standard reference (195 DE/positive and 72 non-DE/negative calls across all three comparisons), we counted the number of TN and TP calls made by each platform at multiple threshold values. The resulting ROC curves (Fig. 4D; Supplemental file “4D-Detailed”) exemplify the effect of a chosen cutoff on the perceived sensitivity and specificity of each microarray platform. The threshold values generating the highest overall number of TP calls for each platform was determined to be optimal and is consistent with the ratio compression of each platform such that the platforms exhibiting greater compression (e.g., Combimatrix) perform better at lower cutoffs than those with lower compression (e.g., Ambion). The number of TP and FP DE and non-DE calls made by each platform is presented in Figure 4E (qPCR bars).

Unexpectedly, some outliers in Figure 4C are miRNAs that correlate poorly with qPCR across all platforms (colored circles), suggesting that the FPs were generated by qPCR (similarly to a recent observation by Ach et al. 2008), rather than consistent errors across platforms incorporating different probe design, hybridization conditions, and labeling chemistries. We therefore repeated the DE analysis with the qPCR data incorporated into the iMLE algorithm. Figure 4E contrasts the number of true and false calls made by each platform at the optimal cutoffs calculated using qPCR either as a reference or integrated into iMLE. Consistently across all platforms, the number of true calls calculated under the iMLE algorithm was greater than those calculated using qPCR as a gold standard. The iMLE (TP/TN) rates are as follows: Agilent, 0.90/0.86; Exiqon, 0.82/0.85; Illumina, 0.87/0.71; Ambion, 0.91/0.91; Combimatrix, 0.59/0.95; Invitrogen, 0.61/0.67; qPCR, 0.83/0.71; and GAsq, 0.85/0.56. Omission of “obsolete” or “not_designed_for_hsa” Exiqon probes resulted in minimal changes to these numbers (± 0.2 in optimal fold-change cutoff and ± 0.04 in TP/TN rates; data not shown). The low sensitivity (TP) of GAsq contradicts the commonly expressed expectation of digital miRNA profiling and was also recently reported in a comparative study using a pool of synthetic RNAs (Willenbrock et al. 2009).

DISCUSSION

We present a comparison of the suitability of six microarray platforms and one NGS technology to detect differential expression of miRNAs. In our hands, Ambion,

Agilent, and Exiqon microarrays ranked highest in the rate of true DE calls. During the course of this study, several changes occurred in the handling protocols and microarray design, some of which are summarized by the manufacturers in the Supplemental file “Manufacturers Comments.” Moreover, NGS and miRNA microarrays are now available from several additional manufacturers (e.g., Affymetrix microarrays, whose performance in comparison to Agilent and Exiqon is currently under evaluation by ABRF) (Web-report 2009). We therefore delineate generic key criteria for the evaluation of current miRNA platforms, including common aspects of microarray technology, such as reproducibility, and aspects particular to miRNAs, such as probe annotation and the utility of qPCR for validation.

Several practical considerations are worthy of mention in miRNA microarray platform selection (Table 1). The choice of single- or dual-channel platform depends on the nature of the biological question investigated, and reliable data were generated by all three single channel platforms and the Ambion dual-channel platform. We found that despite the overall lower signal intensity of cell lines, all platforms were equally applicable to cell line and tissue samples (to be corroborated in additional tissues). The platforms vary widely in their input sample requirement, ranging from 100 ng of total RNA (Agilent) to small RNA-enriched fractions equivalent to ~ 10 μ g total RNA (Ambion and Combimatrix). Thus, despite Ambion’s excellent TP and TN rates, the platform is not suitable for studies where input material is limited. Similarly, Ambion’s performance in detection of DE may be secondary to ease of handling or slide layout in studies with large numbers of samples, or in a high-throughput core facility, for which the labeling and hybridization protocols of Agilent and Combimatrix would be better suited. Platforms also varied in the reproducibility of hybridization, enumerated as CV across replicates (Fig. 1B,C) and consistency of present/absent calls (Fig. 2A). Lower reproducibility might prescribe a larger number of replicate arrays, affecting the experimental design, computational analysis and costing. Cross-hybridization can be estimated by the signal distribution and present calls from mouse and mismatched human probes as a surrogate measure (Fig. 2B,C). Surprisingly, the LNA probes used by Exiqon were among the poorest in discriminating the groups of probes classified using our uniform reannotation, although the contribution of suboptimal LNA spike patterns could not be evaluated. Finally, unique features such as the ability to customize the microarray probe sets for specific applications (Agilent and Combimatrix), or supported array stripping and reuse procedures (Combimatrix), come into play for particular experimental needs.

Periodic changes to miRBase necessitate a reannotation of microarray and qPCR probes prior to analysis. For example, 35 novel miRNAs of each Ambion and Exiqon match recent additions to miRBase. Our arrays, although

acquired within a few weeks of each other, were designed and annotated against different versions of miRBase, resulting in a low number of initially overlapping miRNA identifiers. A substantial fraction of the discrepancies resulting from changes in miRNA nomenclature can be resolved by consulting the tracking files available on miRBase without further computational manipulation. However, changes to the actual sequences of miRBase entries expose potential cross-hybridization between previously unrelated probes and therefore must be identified computationally. Unfortunately, the sequence information provided by the manufacturers is often partial (e.g., miRNA target rather than probe, or probe without proprietary linker). At the two extremes, Combi-matrix provides all probe sequences whereas Exiqon offers only proprietary reannotation against miRBase updates, reserving probe sequence information for users bound by confidentiality agreements. This model restricts the inclusion of sequence information in published research studies. Laboratories with no access to fully exploratory methods (such as deep sequencing) may benefit from microarray platforms that include novel miRNAs (Ambion, Exiqon; annotated by the manufacturers), provided that the underlying probe sequences are disclosed.

High-throughput sequencing of miRNAs is coming into wider use and is unmatched for the discovery and experimental validation of novel or predicted miRNAs. However, library preparation methods seem to have systematic preferential representation of the miRNA complement, resulting in different DE calls (Linsen et al. 2009) and the approach awaits rigorous evaluation. We therefore focused on the differential expression of 204 miRNAs represented by all six microarray platforms as well as detected by sequencing. We observed a low degree of overlap in the DE miRNAs (consistent with Sato et al. 2009), not easily attributable to the strength or weakness of singular platforms. We implemented a novel algorithm (iMLE) integrating partial overlaps of DE calls in the calculation of TP and TN rates. Furthermore, we show that qPCR is not an infallible validation method of miRNA microarray data, especially where the array technology itself incorporates PCR-based amplification (e.g., Illumina). The question of an “industry standard” in miRNA expression awaits further advances in both technology (e.g., deep sequencing) and computation (normalization and DE algorithms). iMLE-based assignment of true values can also potentially help amalgamate other binary datasets, such as peak-calling or miRNA target predictions by different algorithms with no need for a standard reference.

We illustrate the effect of using non-miRNA reference genes for qPCR normalization on the perceived differential expression of tested miRNAs. This effect is pronounced when the overall abundance of miRNAs varies, e.g., in experiments affecting the miRNA processing machinery, or in comparisons involving multiple tissues (such as demonstrated by Sato et al. 2009) or combinations of tissues and cell lines. In such

cases, it is advisable to perform qPCR measurements of numerous miRNAs, including those identified as stably expressed, to obtain a measure of the linear correlation intercept prior to assignment of validated DE values. Alternatively, microarrays and NGS can be used for mutual validation, circumventing the need for external references.

To our knowledge this is the first systematic study scrutinizing the relative performance of miRNA microarrays, NGS, and qPCR across several well-studied biological samples. While our analysis is not intended to serve as a recommendation for any particular platform, we present practical criteria and metrics to evaluate the reproducibility, specificity, and reliability of methods measuring miRNA expression.

MATERIALS AND METHODS

Preparation of total RNA and small-RNA enriched samples

A pool of commercial normal breast tissue (hereafter termed Normal) total RNAs was created from 78 μ g comprising a five-donor pool (BioChain Institute, lot no. A512460), 130 μ g Hm breast total RNA (Ambion AM6952, lot no. 02060262), and 75 μ g MVP human adult breast total RNA (Stratagene 540045-41, lot no. 0870161). The breast cancer cell lines PMC42 (a gift from Michael O'Hare, University College London) (Whitehead et al. 1983, 1984) and MCF7 (from ATCC) (Soule et al. 1973) were cultured in RPMI or DMEM media (Invitrogen), respectively, supplemented with 10% bovine calf serum (Invitrogen). RNA was extracted from subconfluent cultures (estimated 85% density) that were refed with fresh medium 24 h prior to harvesting. In brief, cultures were washed once with cold phosphate-buffered saline (PBS). Upon complete removal of the PBS, cells were lysed directly in 8.4 mL of QIAzol (Qiagen), and total RNA was extracted using 10 miRNeasy columns (Qiagen) according to manufacturer's recommendations.

Several 100 μ g aliquots from each total RNA were further separated into large- and small-RNA enriched fractions (cutoff \sim 200 nt) using the miRNeasy columns and reagents. The yield and quality of the total RNA were monitored by spectrophotometry at 260, 280, and 230 nm, by Agarose gel electrophoresis, and on a Bioanalyzer Eukaryote Total RNA Nano Series II chip (Agilent). RNA integrity number (RIN) values were 9.4 (MCF7), 10.0 (PMC42), and 7.6 (Normal). The yield and quality of the small-RNA enriched fraction (sRef) were monitored by spectrophotometry (as above), urea/polyacrylamide gel electrophoresis (Git et al. 2008), and on a Bioanalyzer Small RNA Series II chip (Agilent). sRef were extracted with a near 100% efficiency, contained predominantly tRNA and small rRNA, and comprised a different but reproducible proportion of the total RNA in each sample: $14 \pm 1\%$ in MCF7, $12.5 \pm 0.5\%$ in PMC42, and $6 \pm 0.2\%$ in normal breast tissue. The miRNA contained within these fractions was <0.5 ng per 10 μ g total RNA (Git et al. 2008; data not shown).

Microarray study design

For single-channel platforms (Agilent, Illumina), each sample was hybridized in quadruplicate (samples are termed M, MCF7; P,

PMC42; and N, Normal throughout the text). For dual-channel platforms, a balanced-dye design was employed in which quadruplicate hybridizations were set up in the following combinations: Cy3-MCF7 with Cy5-PMC42 (sample MP), Cy3-PMC42 with Cy5-Normal (sample PN), and Cy3-Normal with Cy5-MCF7 (sample NM).

The hybridizations for each quadruplicate were carried out on two different days. Where possible replicates that were labeled side-by-side were hybridized on different days, and those labeled on different days were hybridized side by side. For microarray platforms requiring near immediate application of the labeled samples (Combimatrix, Exiqon), independent labeling reactions were carried out on the day of the hybridization. For Ambion assays, two independent sets of duplicate dried polyadenylation reactions were frozen for <48 h, and labeling of individual replicates was completed immediately prior to hybridization.

RNA labeling and microarray hybridization

RNA input and labeling kits were chosen and used according to the recommendations of each microarray manufacturer (Kreatech labeling for Combimatrix arrays and the manufacturers' labeling kits for others). Arrays were hybridized for 16–20 h in an Agilent G2545A hybridization oven and washed according to the manufacturer's instructions. To minimize bias due to seasonal changes in ultraviolet light and ambient ozone, we completed all in-house experimental work at one location over a span of 6 wk.

Agilent

One hundred nanograms of total RNA samples was dephosphorylated, 3' end-labeled with Cy3-pCp, purified on Micro Bio-Spin columns, dried, and hybridized using miRNA Microarray System labeling kit and arrays (Agilent) (Wang et al. 2007).

Ambion

sRef samples equivalent to 10 µg total RNA were polyadenylated, purified, dried to completion, coupled to Cy3 or Cy5 amine-reactive dyes (GE Healthcare), purified, dried, and hybridized using mirVana miRNA Labeling and Bioarrays Version 2 (Ambion) (Shingara et al. 2005).

Combimatrix

sRef samples equivalent to 10 µg total RNA were coupled to Cy3- or Cy5-ULS reagent using ULS Small RNA Labeling kit (Kreatech) and hybridized to MicroRNA 4X2K Microarrays (Combimatrix).

Exiqon

One microgram total RNA samples was dephosphorylated, Hy3- or Hy5- end-labeled, and hybridized using miRCURY LNA microRNA Array Power Labeling kit and microarray kit (Exiqon).

Illumina

Two hundred nanograms of total RNA samples were processed by Illumina using a Sentrix Array Matrix 96-well MicroRNA Expression Profiling Assay v1 (Chen et al. 2008). In brief, samples are polyadenylated and reverse-transcribed, and the cDNA is hybridized to a specific primer pool and extended to incorporate address

tags and universal sequences. PCR-amplified samples are then hybridized to address-coded beads on a solid support.

Invitrogen

One microgram of total RNA samples was polyadenylated, 3' splint-ligated to Cy3- or Cy5-labeled oligonucleotides, and hybridized using NCode Rapid miRNA Labeling System and NCode Multi-Species miRNA Microarray v2 (Invitrogen).

Microarray scanning and feature extraction

Illumina bead-based arrays were processed at the manufacturer's facility in San Diego, California. In brief, arrays were scanned on a BeadScan instrument, and fluorescence intensities were extracted and summarized using the BeadStudio software (Illumina), resulting in a set of summarized fluorescence measurements (Supplemental file "MPN_miRNA_Illumina"). Agilent, Ambion, Exiqon, and Invitrogen arrays were scanned on a G2505B Microarray Scanner (Agilent Technologies), and Combimatrix arrays were scanned on the InnoScan700 (Innopsys). Feature recognition and alignment of all in-house scanned images were carried out using GenePix Pro 6.1 and, where necessary, adjusted manually by the same operator. To minimize variation in alignment correction, arrays from each platform were processed in a single session. Data from Agilent, Ambion, Combimatrix, Exiqon, and Invitrogen arrays have been deposited in ArrayExpress (<http://www.ebi.ac.uk/microarray-as/ae/>; accession E-MTAB-96).

Microarray normalization and processing

Data analyses were carried out within the R statistical computing framework version 2.8.0 (<http://www.R-project.org/>) (R Development Core Team 2008). Following quality control assessment, two out of the 60 arrays hybridized in-house were excluded (see Fig. 1B) due to either low intensity (Exiqon, sample M) or array-specific artifacts (Combimatrix, sample PN). The overall Cy5 intensities in Exiqon arrays were too low for reliable analysis, and the data from the Cy3 channel was treated as a single-channel assay.

The limma package (Smyth 2005) was used for microarray processing. Different methods for background correction were tested for all platforms except Illumina (*none*, *subtract*, *half*, *minimum*, *movingmin*, *normexp*) and normalization (*none*, *vsu*, *quantile*), depending on whether the platforms were used for a single- or dual-channel assay. Ultimately, *normexp* was chosen due to its superior performance in correcting spatial artifacts, maximizing the uniformity of foreground and background signal, and minimizing the variability within and between arrays. All platforms were background-corrected using *normexp*, except for Combimatrix, where *minimum* was used due to constraints specific to the array layout. Dual-channel platforms (Ambion, Combimatrix, Invitrogen) were normalized using *loess* spatial correction within arrays, and single-channel platforms (Agilent, Exiqon) were quantile normalized between arrays. Spike-in controls were not used for normalization purposes as they were only available for some of the platforms, and where present were too few to be reliably utilized.

SNR was calculated using the SSDR method (He and Zhou 2008), $\mu_i/(\sigma_{iF} + \sigma_{iB})$ (where μ equals spot intensity; σ , pixel standard deviation; i, spot; F, foreground; and B, background).

Microarray probe reannotation

All probe sequences were mapped to mature human and mouse miRNA sequences from miRBase version 12 (Griffiths-Jones et al. 2008) using WU-BLAST (Lopez et al. 2003). Ungapped alignment was performed, using word length shorter than the default when necessary. For “long probe” platforms (Ambion, Invitrogen, Combimatrix, Illumina), all perfect match hits with length greater than 15 were retained and filtered as described below. For “short probe” platforms (Agilent, Exiqon), probes with length greater than 15 were treated similar to the “long probes” platforms, whereas for short probes only perfect match hits with alignment length equal to probe length were considered. Where alignments ≥ 20 bases were found, shorter alignments were discarded. For alignments < 20 bases, the longest was assigned to the given probe, or multiple miRNAs in the case of matches of equal length. For alignments ≥ 20 , there were occasionally several possible miRNAs targets, and these were all assigned to the probe to account for potential cross-hybridization. A complete list of reannotated probes can be found in Supplemental file “Reannotation.”

In cases where a probe sequence aligned to both a human and mouse miRNA, targets were assigned to the probe under the following priorities: human perfect match > human with one mismatch > mouse miRNAs. Probes were finally grouped into the following categories: PosControl, NegControl, and Novel: positive or negative controls and putative novel miRNAs, respectively, as defined by array manufacturer; Empty indicates spots with no printed probes; Human and MM_human, probes targeting human miRNAs with perfect complementarity or with a single internal mismatch, respectively; Mouse, probes targeting mouse, but not human, miRNAs; Obsolete, probes that were designed to target miRNAs but do not map to targets in the current version of miRBase; and Miscellaneous, probes outside the aforementioned categories, such as probes targeting miRNAs from other species, spike-in controls, and all unidentifiable probes.

We examined the signal intensities across probes of different lengths and GC content. Some variation was observed (data not shown), but since the binding kinetics for individual platforms are affected by numerous factors, we did not attempt to correct for these in the analysis.

Putative polycistronic miRNAs were defined as sets of miRNAs sharing a genomic locus with no more than 500 bases between any two adjacent miRNAs, and were obtained via the Clusters interface of miRGen (Megraw et al. 2007).

Assignment of microarray present and absent spots

Spots were called as “absent,” “marginally present,” or “present” using a modified version of the R package “panp” (Warren et al. 2007). A probability distribution of signal intensities from empty and negative control spots was calculated, and the cumulative distribution function (CDF) generated. Each spot was called as present or absent based on expression value cutoffs defined from the survivor distribution (1-CDF) for each individual array, using *P*-values of 0.05 (present) and 0.1 (marginally present). For dual-channel arrays (Ambion, Combimatrix, and Invitrogen), each channel was treated separately, and the percentage of present calls for each miRNA was taken across both the Cy3- and Cy5-labeled fluorescence data.

Identification of differentially expressed miRNAs

For single-channel platforms, *M*-values were calculated based on the individual Cy3 data from each sample (hereafter also referred to as *M*-values). Ratio compression was taken as the slope of the linear least-squares regression of microarray versus qPCR across all three biological comparisons. All *M*- and *P*-values are available in Supplemental file “M_pValue_204probes.” The empirical Bayes moderated *f*-statistics implemented in the R package limma was used. Differentially expressed genes were identified using the limma nestedF procedure, applying a significance threshold of 0.05 in combination with Benjamini–Hochberg false-discovery rate control and unless otherwise specified, a minimal cutoff of 2. Where multiple probes targeting the same miRNA did not agree, one of two approaches was chosen for clarity of presentation: For Figure 4D, we have assigned the corresponding miRNA with an “NA” value, while for Figure 4E, the miRNA was assigned with the value of the probe that showed differential expression, as long as the two calls were not contradictory (up- and down-regulated) in which case the miRNA was assigned “NA.”

Next-generation sequencing

sRef samples equivalent to 2 μ g total RNA were ligated to a preadenylylated 3' adapter v1.5 (5'-rApp-[desoxy]ATCTCGTATG CCGTCTTCTGCTTG-[didesoxy]ddC-3'; Illumina or Dharmacon) in 1 \times T4 RNL2 truncated reaction buffer (NEB), 10 mM MgCl₂ (Ambion), 20 units of RNaseOUT (Invitrogen), and 300 units truncated T4 RNA ligase 2 for 1 h at 22°C. The reactions were then supplemented with 12.5 nmol 5' adapter (all RNA; GUUCA GAGUUCUACAGUCCGACGAUC; Dharmacon), 1 mM ATP (Ambion), and 20 units of T4 RNA ligase (NEB) and the second ligation allowed to proceed for 6 h at 20°C. The double-ligation products were reverse-transcribed by SuperScriptII reverse transcriptase (Invitrogen) in the presence of primer GX1 (all desoxy; CAAGCAGAAGACGGCATACGA; Sigma) following manufacturer's instructions. The cDNA was PCR-amplified by Phusion DNA Polymerase with Primers GX1 and GX2 (all DNA; AATG ATACGGCGACACCGACAGGTTTCAGAGTTCTACAGTCCGA; Sigma) for 19 cycles of [10 sec at 98°C, 30 sec at 60°C, and 15 sec at 72°C]. The amplification products were separated on a Novex 6% TBE gel (Invitrogen), and the 90–100 base-pair bands were excised, eluted into 0.3 M NaCl, and ethanol precipitated. Following quality control on a Bioanalyzer 1000 DNA chip (Agilent), the purified DNA fragments were used directly for two independent repeats of sequencing via 36 alternating cycles of enzymatic synthesis and optical interrogation using the Illumina Cluster Station and GAII Genome Analyzer following manufacturer's protocols. Sequencing reads were extracted from the image files generated by Genome Analyzer II using the GAPIipeline software, version 1.4 (Illumina).

NGS data analysis

3' adapters were trimmed from sequencing reads using an in-house script (available upon request). Reads of length < 15 nt after adapter trimming and comprising more than 50% polyA stretches were excluded from further analyses. The remaining reads were mapped to known mature miRNAs (miRBase version 12) using the “ssaha2” program (Ning et al. 2001), where 100% identity between reads and known miRNAs sequences was required. miRNAs with an aggregate count of less than 10 in all samples

were eliminated (see Supplemental file “GA_Read_Counts”); then the total read count for each lane was scaled relative to the library size (total number of reads that mapped to known miRNAs). Read counts of technical replicates were then merged, and \log_2 (fold-change) values were calculated for each miRNA. *P*-values were subsequently calculated using a binomial approximation to Fisher’s exact test for each miRNA.

Real-time RT-PCR (qPCR)

For SYBR green-based assays, sRef samples equivalent to 10 μ g total RNA were polyadenylated, reverse-transcribed using a tagged and anchored oligo-dT primer, and then amplified using a gene-specific forward primer and universal reverse primer (see Supplemental file “qPCR Primers”) in the presence of SYBR green as described by Git et al. (2008). For TaqMan assays (ABI), total RNA samples were reverse-transcribed using a pool of gene-specific primers and amplified using individual gene-specific assays. All RT reactions were performed with three different RNA inputs, and all PCR reactions were carried out in triplicate. RNU48 and 5S rRNA were used as non-miRNA reference genes for TaqMan and SYBR green qPCR, respectively. The measured Ct values were M:11.24, P:11.63, N:11.70 (RNU48) and M:16.20, P:16.50, N:16.27 (5S rRNA), and the magnitude of the variance did not warrant $\Delta\Delta$ Ct normalization. Where miRNAs were tested by both methods, the average correlation was 0.94.

iMLE algorithm

The input for the algorithm is a table of discrete DE calls (up-regulated, +1; not DE, 0; down-regulated, -1) for each miRNA/comparison combination (rows) made by each experimental platform at a particular threshold value with *P*-value <0.05 (columns). An initial “truth” value was assigned for each row according to the majority of calls. A matrix (i,j) was then generated for each platform, representing the proportion of cases where the assay called various *j* for each Truth *i*, e.g., $P(-1,-1) + P(-1,0) + P(-1,+1) = 1$. Subsequently the algorithm reiterated two steps until Truth values converged: (1) selected for each row the Truth (-1/0/+1) with the highest maximal likelihood estimate [MLE, defined as the product of all platform probabilities to have given this Truth call under the existing (i,j) parameters], followed by (2) a recalculation of the platform matrices.

To determine the optimal cutoffs for each platform, the iMLE was performed in an iterative fashion, where cutoffs were fixed for all but one tested platform at a time, for which a series of discrete cutoffs was tested, and the cutoff that yielded the highest number of correct calls was fixed as a temporary optimum. This was repeated across all platforms until each platform cutoff converged to a stable value. The following measures were then extracted from the platform matrices: TP [the average of (1,1) and (-1,-1)]; TN (0,0); FP [average of (0,1) and (0, -1)]; FN [average of (1,0) and (-1,0)]; reverse [average of (1, -1) and (-1,1)].

An outline of the algorithm is included in Supplemental file “iMLE Algorithm,” and the code for the implementation of the algorithm is available from the authors upon request.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at <http://www.rnajournal.org>.

ACKNOWLEDGMENTS

We thank Yoav Git for advice in algorithm design; and Sarah Moffatt, Nick Matthews, and Rory Stark for help with sequencing the small RNA libraries. C.C., J.H. and A.G. conceived and co-ordinated the study; A.G. extracted and labeled the RNA, extracted feature intensities after scanning and participated in the analysis; M.O. carried out the hybridizations and scans; C.K. prepared the small RNA libraries; H.D. and M.S.-D. analyzed the data with advice from P.B.; and A.G., H.D., and M.S.-D. drafted the manuscript; which was approved by all authors. This work was supported by the University of Cambridge, Cancer Research UK, Hutchison Whampoa Limited, the European Molecular Biology Laboratory, and the Swiss National Science Foundation.

Received October 1, 2009; accepted February 4, 2010.

REFERENCES

- Ach RA, Wang H, Curry B. 2008. Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnol* **8**: 69.
- Barbato C, Giorgi C, Catalanotto C, Cogoni C. 2008. Thinking about RNA? MicroRNAs in the brain. *Mamm Genome* **19**: 541–551.
- Baumbusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, Bruhn L, Gunderson K, Naume B, Kristensen VN, Liestol K, et al. 2008. Comparison of the Agilent, ROMA/NimbleGen, and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9**: 379.
- Bissels U, Wild S, Tomiuk S, Holste A, Hafner M, Tuschl T, Bosio A. 2009. Absolute quantification of microRNAs by using a universal reference. *RNA* **15**: 2375–2384.
- Bueno MJ, de Castro IP, Malumbres M. 2008. Control of cell proliferation pathways by microRNAs. *Cell Cycle* **7**: 3143–3148.
- Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, et al. 2005. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* **33**: e179. doi: 10.1093/nar/gni178.
- Chen J, Lozach J, Garcia EW, Barnes B, Luo S, Mikoulitch I, Zhou L, Schroth G, Fan JB. 2008. Highly sensitive and specific microRNA expression profiling using BeadArray technology. *Nucleic Acids Res* **36**: e87. doi: 10.1093/nar/gkn387.
- Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92–105.
- Git A, Spiteri I, Blenkiron C, Dunning MJ, Pole JC, Chin SF, Wang Y, Smith J, Livesey FJ, Caldas C. 2008. PMC42, a breast progenitor cancer cell line, has normal-like mRNA and microRNA transcriptomes. *Breast Cancer Res* **10**: R54.
- Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: Tools for microRNA genomics. *Nucleic Acids Res* **36**: D154–D158.
- Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**: 3–12.
- He Z, Zhou J. 2008. Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Appl Environ Microbiol* **74**: 2957–2966.
- He X, Eberhart JK, Postlethwait JH. 2009. MicroRNAs and micro-managing the skeleton in disease, development, and evolution. *J Cell Mol Med* **13**: 606–618.
- Hennessy E, O’Driscoll L. 2008. Molecular medicine of microRNAs: Structure, function, and implications for diabetes. *Expert Rev Mol Med* **10**: e24. doi: 10.1017/S1462399408000781.
- Hua YJ, Tu K, Tang ZY, Li YX, Xiao HS. 2008. Comparison of normalization methods with microRNA microarray. *Genomics* **92**: 122–128.

- Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. 2009. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* **37**: D98–D104.
- Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, Ghosh J, Brizuela L, Carroll JS, Brown M, Flicek P, et al. 2008. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* **18**: 393–403.
- Jovanovic M, Hengartner MO. 2006. miRNAs and apoptosis: RNAs to die for. *Oncogene* **25**: 6176–6187.
- Krutzfeldt J, Stoffel M. 2006. MicroRNAs: A new class of regulatory genes affecting metabolism. *Cell Metab* **4**: 9–12.
- Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**: 853–858.
- Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**: 1401–1414.
- Lee EJ, Baek M, Gusev Y, Brackett DJ, Nuovo GJ, Schmittgen TD. 2008. Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. *RNA* **14**: 35–42.
- Li W, Ruan K. 2009. MicroRNA detection by microarray. *Anal Bioanal Chem* **394**: 1117–1124.
- Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**: 474–476.
- Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W. 2003. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res* **31**: 3795–3798.
- Medina PP, Slack FJ. 2008. microRNAs and cancer: An overview. *Cell Cycle* **7**: 2485–2492.
- Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. 2007. miRGen: A database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res* **35**: D149–D155.
- Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res* **11**: 1725–1729.
- Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, et al. 2004. Identification of virus-encoded microRNAs. *Science* **304**: 734–736.
- Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K. 2009. Impact of normalization on miRNA microarray expression profiling. *RNA* **15**: 493–501.
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roush S, Slack FJ. 2008. The let-7 family of microRNAs. *Trends Cell Biol* **18**: 505–516.
- Sato F, Tsuchiya S, Terasawa K, Tsujimoto G. 2009. Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. *PLoS One* **4**: e5540. doi: 10.1371/journal.pone.0005540.
- Sheng Y, Engstrom PG, Lenhard B. 2007. Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PLoS One* **2**: e946. doi: 10.1371/journal.pone.0000946.
- Shi R, Chiang VL. 2005. Facile means for quantifying microRNA expression by real-time PCR. *Biotechniques* **39**: 519–525.
- Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**: 1151–1161.
- Shingara J, Keiger K, Shelton J, Laosinchai-Wolf W, Powers P, Conrad R, Brown D, Labourier E. 2005. An optimized isolation and labeling platform for accurate microRNA expression profiling. *RNA* **11**: 1461–1470.
- Smyth GK. 2005. Limma: Linear models for microarray data. In *Bioinformatics and computational biology solutions using R and bioconductor* (ed. R Gentleman, et al.), pp. 397–420. Springer, New York.
- Soule HD, Vazquez J, Long A, Albert S, Brennan M. 1973. A human cell line from a pleural effusion derived from a breast carcinoma. *J Natl Cancer Inst* **51**: 1409–1416.
- Stefani G, Slack FJ. 2008. Small noncoding RNAs in animal development. *Nat Rev Mol Cell Biol* **9**: 219–230.
- van den Berg A, Mols J, Han J. 2008. RISC-target interaction: Cleavage and translational suppression. *Biochim Biophys Acta* **1779**: 668–677.
- Wang H, Ach RA, Curry B. 2007. Direct and sensitive miRNA profiling from low-input total RNA. *RNA* **13**: 151–159.
- Warren P, Taylor D, Martini PGV, Jackson J, Bienkowska J. 2007. PANP—a new method of gene detection on oligonucleotide expression arrays. In *7th IEEE International Conference on Bioinformatics and Bioengineering, 2007*, pp. 108–115, Boston, MA.
- Web-report. 2009. ABRF study explores microRNA array platforms. *GenomeWeb daily news*. <http://www.genomeweb.com/print/911534?page=show>.
- Whitehead RH, Bertoncello I, Webber LM, Pedersen JS. 1983. A new human breast carcinoma cell line (PMC42) with stem cell characteristics. I. Morphologic characterization. *J Natl Cancer Inst* **70**: 649–661.
- Whitehead RH, Quirk SJ, Vitali AA, Funder JW, Sutherland RL, Murphy LC. 1984. A new human breast carcinoma cell line (PMC42) with stem cell characteristics. III. Hormone receptor status and responsiveness. *J Natl Cancer Inst* **73**: 643–648.
- Willenbrock H, Salomon J, Sokilde R, Barken KB, Hansen TN, Nielsen FC, Moller S, Litman T. 2009. Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing. *RNA* **15**: 2028–2034.
- Yin JQ, Zhao RC, Morris KV. 2008. Profiling microRNA expression with microarrays. *Trends Biotechnol* **26**: 70–76.

RESEARCH ARTICLE

Open Access

The cost of reducing starting RNA quantity for Illumina BeadArrays: A bead-level dilution experiment

Andy G Lynch^{1,2*}, James Hadfield², Mark J Dunning², Michelle Osborne², Natalie P Thorne³, Simon Tavaré^{1,2}

Abstract

Background: The demands of microarray expression technologies for quantities of RNA place a limit on the questions they can address. As a consequence, the RNA requirements have reduced over time as technologies have improved. In this paper we investigate the costs of reducing the starting quantity of RNA for the Illumina BeadArray platform. This we do via a dilution data set generated from two reference RNA sources that have become the standard for investigations into microarray and sequencing technologies.

Results: We find that the starting quantity of RNA has an effect on observed intensities despite the fact that the quantity of cRNA being hybridized remains constant. We see a loss of sensitivity when using lower quantities of RNA, but no great rise in the false positive rate. Even with 10 ng of starting RNA, the positive results are reliable although many differentially expressed genes are missed. We see that there is some scope for combining data from samples that have contributed differing quantities of RNA, but note also that sample sizes should increase to compensate for the loss of signal-to-noise when using low quantities of starting RNA.

Conclusions: The BeadArray platform maintains a low false discovery rate even when small amounts of starting RNA are used. In contrast, the sensitivity of the platform drops off noticeably over the same range. Thus, those conducting experiments should not opt for low quantities of starting RNA without consideration of the costs of doing so. The implications for experimental design, and the integration of data from different starting quantities, are complex.

Background

Gene expression microarrays have become a routine analysis tool; from their introduction [1] to recent headline publications [2-4] their widening use has been primarily down to better understanding of how to design [5,6], use and analyse [7,8] microarray experiments. An important, if somewhat forgotten, design issue has been the amount of starting material needed to produce high quality microarray data. Ten years ago, around 10 μ g of total RNA was required and even three years ago many labelling protocols required 1 μ g. The introduction of Illumina BeadChips with a standard labelling reaction requiring only 250 ng of total RNA made analysis of some previously unconsidered sample types possible;

e.g. limited clinical samples or samples requiring considerable microdissection and pooling.

Whilst many researchers continue to push the limits of starting materials [9], development of robust standard labelling protocols has further decreased the amount of RNA required for microarrays. Until recently 250 ng of starting mRNA was recommended for the Illumina BeadArray platform. Now 50 ng to 100 ng is suggested http://www.illumina.com/technology/direct_hybridization_assay.ilmn. If one can indeed use so little starting material then this is of tremendous importance in terms of the scope of experiments that become possible. However, there is a wealth of literature that is based upon 250 ng, and it is important that future results are comparable to those obtained previously. One small comparison has previously been made [10]. This study found that reproducible signal was obtainable from as little as

* Correspondence: andy.lynch@cancer.org.uk

¹Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

Full list of author information is available at the end of the article

25 ng, but the study was not large enough to quantify the costs of such an approach.

Microarray dilution experiments [11,12], where two samples are mixed together in a number of differing (known) ratios and those mixtures hybridized to arrays, have proven to be valuable tools for the comparison and investigation of microarray platforms, most notably in the MAQC project [13]. We employ a nine-level dilution design to investigate the effect of changing the quantity of starting mRNA on the performance of Illumina BeadArrays. We consider the previously recommended level of 250 ng, the current recommended levels of 100 ng and 50 ng and one other (10 ng).

Here, we examine the costs and consequences of reducing the amount of starting RNA, with consideration for the issues of experimental design and meta-analysis, while also providing a unique bead-level dilution experiment to serve as a public resource to the Illumina-using community. We use the Illumina HumanWG-6 V3 BeadArray, analysed at the bead-level as we have previously recommended [14]. One of the benefits of using the bead-level data is that we can analyse separately the two array-sections assigned to any one sample, thus allowing inferences to be made about the more flexible HT12 BeadArrays also. In addition to our purposes, we are creating a unique public resource, and have designed our experiment to be generally useful to the community.

Methods

Experimental Design

Six samples can be hybridized to the Illumina HumanWG-6 V3 chip, each sample on two array-sections of approximately 1,000,000 beads that are distributed amongst approximately 50,000 bead-types. We treat the two sections as separate arrays for the purposes of analysis, due to previously observed inter-section differences [14,15]. This also has the effect of making our results comparable to those one might expect from the Illumina HT-12 array which takes 12 samples, allocated one section each.

We have used two reference RNA samples, previously employed in the MAQC study [13], which have subsequently become a standard for microarray [16] and next-generation sequencing [17] studies. These are the Stratagene Universal Human Reference RNA (hereafter "UHRR"), and the Ambion Human Brain Reference RNA (hereafter "Brain"). Nine levels of mixture, including the four employed in the first MAQC study, were then created. These are 100:0, 99:1, 95:5, 90:10, 75:25, 50:50, 25:75, 10:90 and 0:100, where mixtures are presented as UHRR:Brain.

These nine levels allow for investigation of broad trends, and for the detection of subtle differences. Combined with the four levels of starting material that we

are investigating, this leads to 36 samples to be arranged across six Illumina HumanWG6 V3 BeadChips. Clearly it would not be desirable to confound levels of starting material with BeadChips as we would be unable to distinguish our comparison of interest from technical variation. However it is desirable that our data resemble data from a 'real-world' experiment else they have no external validity and, in general, experiments are conducted on BeadChips using only one level of starting material.

Our design was chosen to address this tension between internal and external validity. Each BeadChip was run with samples from two starting quantities of RNA (three samples from both chosen starting quantities), and each possible combination of the two starting quantities was run once and only once amongst the six BeadChips. Full details of the design are given in Section 1 of Additional File 1.

Laboratory methods

Stock UHRR tubes were prepared following manufacturer's recommendation and pooled to create a stock of 1 mg/ml; Brain RNA was received at 1 mg/ml. The quality was checked using the Agilent Bioanalyser. The RNA was accurately diluted to a working stock of 100 ng/ μ l and the dilution series was created to the specifications given above. The minimum pipetting volume used was 10 μ l.

The Illumina TotalPrep-96 Kit (4397949) was used to process the samples using the range of input concentrations in question. For the 50 ng and 10 ng input quantities a 1:10 dilution of working RNA was used. Quality and quantity of the cRNA was checked before proceeding with hybridisation to Human WG-6 V3 BeadArray. The Illumina WGGX DirectHyb Assay Guide (11286331 RevA) protocol was followed for hybridisation, washing and scanning of the BeadArray, with the scanner set to return bead-level data (Additional File 1, Section 2). Quality assessment was achieved via examination of metrics files (Additional File 1, Section 3), agreement with previous MAQC data sets (Additional File 1, Section 4), and performance of housekeeping controls (Additional File 1, Section 5).

Preprocessing and statistical analysis

Illumina BeadScan files were processed and analysed using the *beadarray* package [18] from Bioconductor. Arrays were pre-processed on the log₂-scale on a per-array-section basis. BASH [19] was used to remove high-frequency spatial artefacts, followed by outlier removal (outliers being defined as observations more than three median absolute deviations from the median), and expression detection score calculation. The detection score is a standard measure for Illumina

expression experiments, and can be viewed as an empirical estimate of the p-value for the null hypothesis that there is no expression. Between-array-section quantile-normalization was performed within each starting material level, and a non-linear regression model fitted across dilution levels within each starting RNA level.

Our approach demands reporting of raw, bead-level, Illumina data, which exceeds the MIAME requirements. As popular repositories are not designed for the storage of raw (bead-level) data from random arrays, the files are available to download from our website at <http://www.compbio.group.cam.ac.uk/Resources/Dilution/Dilution.html>.

Statistical model

A previously proposed [20] non-linear model was used as the theoretical model for the dilution curve:

$$E_{mrp} = \log_2(c_m U_{rp} + (1 - c_m) B_{rp}) + \epsilon_{mrp} \quad (1)$$

where E_{mrp} is the observed (normalized) log-intensity for probe p at starting RNA quantity r in mixture level m , c_m is the proportion of the mixture that is UHRR, U_{rp} is the intensity associated with probe p at starting RNA level r in the UHRR sample, and B_{rp} is similarly defined for the Brain sample. The ϵ_{mrp} are independent measurement errors with mean zero and standard deviation σ_{rp} .

This model implicitly assumes a linear relationship between quantity of RNA and measured intensity. This assumption is known not to hold over the full range of observed intensities for microarrays [21], and specifically for Illumina BeadArrays [14]. While some models allow for non-linearity [22], they do not relate it to the known physico-chemical causes. To do so would be difficult and, in any case, would not obviously be advantageous in our situation.

The model can be rewritten in terms of $\Delta_{rp} = U_{rp} - B_{rp}$,

$$E_{mrp} = \log_2(c_m \Delta_{rp} + B_{rp}) + \epsilon_{mrp} \quad (2)$$

and we fit this model in R using the `nls()` function, weighting each observation by the number of beads that contributed to the observation. Under this formulation, it is clear that the test of $\Delta = 0$ from the `summary.nls()` function in R provides an approximate and quick test of a difference in log-intensities.

Restricting the analysis-group of bead-types

We have re-annotated the bead-types on the array [23], and have identified 23, 562 “perfect” bead-types (using the August 2009 annotation). These are bead-types that have a full 50 mer match to a reliable transcript, and do not possess additional undesirable properties (e.g. mapping to transcripts masked by repeat regions, having a non-unique transcriptomic match, mapping to transcripts that do not align well to the reference genome, etc.). Additionally, we define an ‘analysis-group’ of bead-types as a subset of these perfect bead-types that possess two further properties: 1) That their GC content is conducive to hybridization (i.e. in the range of 20-35 bases), which excludes a further 506 bead-types, and 2) That they occur at least six times on each array-section (see Additional File 1, Section 6). All analyses will be restricted to this ‘analysis-group’ unless otherwise stated.

Results

Numbers of beads

The random assembly of Illumina arrays is often a virtue, but prevents the conduct of true replicate experiments. In particular, the number of usable beads on each array can vary, and will influence performance. There are a number of reasons why disparities emerge. Not all beads are decoded by Illumina when the array is manufactured, (which alone leads to the 10 ng experiment having approximately 80, 000 beads more per array-section than the 100 ng experiment). Further beads are ‘lost’ due to spatial artefacts and to beads being classed as outliers during summarization. The numbers in our experiment are given in Table 1.

It has been observed previously that spatial artefacts can be associated with nearby regions where beads are non-decoded [24], so it may not be coincidental that the

Table 1 Numbers of beads

Quantity of starting RNA:	250 ng	100 ng	50 ng	10 ng
Total decoded	18,801,235	17,835,076	17,926,750	19,274,434
Removed by BASH	200,459	408,721	284,088	50,449
Removed in summarization	651,323	614,495	603,619	582,345
Remaining	17,949,453	16,811,860	17,039,043	18,641,640
In analysis-group bead-types	7,963,638	7,475,940	7,563,440	8,248,259

Summing across all array-sections in the four experiments, we list the total numbers of beads (as decoded by Illumina), the numbers we remove as being in spatial artefacts using BASH [19], those removed as being outliers in the summarization, the remainders, and the numbers remaining that lie in the analysis-group of bead-types.

experiment with the greatest number of beads loses fewest to spatial artefacts. The differing numbers of beads may cause concern, although it should be noted that the median number of replicates for a bead-type only varies from 21 for the 10 ng experiment to a still very healthy 19 for the 100 ng experiment. The lack of monotonicity is also helpful; the trends that we show do not correlate with the total bead-numbers, suggesting that these numbers are not driving the results. Whilst we take 250 ng of starting RNA as our gold standard for comparison, we can also gain reassurance through comparisons to the 100 ng experiment which contained fewest beads.

As noted above, we restrict analyses to an analysis-group containing only 'perfect' bead-types, with desirable GC composition and at least six beads on each array-section. This reduces the number of bead-types considered to 21, 627. This also has a marginal effect on improving the balance between experiments in terms of the numbers of beads analysed.

Detection of expression

In Table 2 is presented a summary of agreement between experiments for the detection of expression (using a significance level of < 0.01) for the analysis-group (see also Additional File 1, Section 7). If no bead-types were truly expressed, we would expect to see 3, 579 apparently showing expression in at least one array-section and nine showing expression in both UHRR and brain. Even acknowledging this, we see that a substantial number of the analysis-group show expression above negative-control levels.

Naturally, any bead-type that shows expression in both Brain and UHRR should show expression in all mixtures of those two samples, and we see that the proportion of bead-types satisfying the former that are also returned by the latter exceeds 80% for the 250 ng, 100 ng and 50 ng experiment but reduces to below 70% for the 10 ng experiment. Agreement between experiments is reported in Table 3, and is encouraging. Performance in terms of sensitivity while not perfect at 100 ng only decreases dramatically when we reach 10 ng, but here still returns 2/3 of the bead-types that were detected in both UHRR and Brain using 250 ng. Notably, the false discovery rate is fairly constant, staying below 10% even

at 10 ng. Thus while one will detect expression in fewer bead-types using less starting RNA, the validity of that which is detected is preserved.

Expression of control bead-types

The detection p-values for expression depend on the performance of negative control bead-types for their calculation. This platform has 759 negative control bead-types, which should have no match to the human transcriptome. Due to the nature of the calculation, at least seven (1%) of these will themselves apparently detect significant expression. Table 4 summarizes the numbers seen in our experiments. We see markedly more than seven negative control bead-types being called as 'detected', and far more than would be expected by chance being consistently called as detected.

Such observations could have explanation other than the bead-types showing specific signal. For instance, thermodynamic variation could lead to some negative control bead-types regularly being called as 'detected', but evidence of differential expression is harder to explain. Using Benjamini-Hochberg control for false discovery rate, there are still three negative control bead-types for the 250 ng starting material experiment (two for 100 ng, eight for 50 ng, and four for 10 ng) that show differential expression. The greatest evidence of a negative control showing specific hybridization is for bead-type ILMN_1343923 (Additional File 1, Section 8).

The amount of starting RNA varies between experiments, but the amount of cRNA used is the same in every case, so there is no reason to anticipate overall changes in intensity levels. However, the intensity levels change for both the housekeeping bead-types (bead-types that target genes *EEF1A1*, *GAPDH*, *TXN*, *ACTB*, *TUBB2A*, *RPS9*, *UBC*) and the negative control bead-types, suggesting that the levels of non-specific hybridization vary according to the amount of starting material (Table 4, Figure 1).

The log-intensity levels for the housekeeping control bead-types decrease at a greater rate than those for the negative control bead-types (except when saturation effects are apparent). Thus the log-fold-change in intensities from housekeeping gene to negative control (a measure of signal to noise) decreases with the

Table 2 Expression detected

Quantity of starting RNA:	250 ng	100 ng	50 ng	10 ng
...at least one array-section	15,880	15,597	15,691	14,090
...both UHRR and Brain	11,992	11,248	10,965	8,775
...all array-sections	9,964	9,178	8,996	5,975
mean number of array-sections	6.94	6.59	6.47	4.95

For each of the four experiments, we report the number of analysis-group bead-types for which expression was detected in at least one of the 18 array-sections, at least one of the two 100% UHRR array-sections and at least one of the two 100% Brain array-sections, and in all 18 array-sections. Additionally for the analysis-group bead-types, we report the mean number of array-sections out of 18 in which expression is detected.

Table 3 Consistency in expression detection between quantities of starting RNA

test experiment	reference experiment		
	250 ng	100 ng	50 ng
10 ng	0.67/0.09	0.71/0.09	0.73/0.09
50 ng	0.86/0.06	0.89/0.09	
100 ng	0.86/0.08		

For the numbers of bead-types with detected expression in both 100% Brain and 100% UHRR reporting "X/Y" where X is the proportion of bead-types reported for the reference experiment also reported for the test experiment (e.g. a measure of sensitivity), and Y is the proportion of bead-types reported by the test experiment that were not reported by the reference experiment (e.g. FDR). So for example, taking 250 ng as a gold-standard, for this detection measure the 100 ng experiment has 86% sensitivity and an FDR of 0.08.

amount of starting RNA. This change in performance is apparent even at 100 ng levels of starting material. Other control bead-types on the Illumina BeadArray platform are not sample dependent, and do not vary considerably between starting quantities of RNA.

Magnitude of expression

Figure 2 shows MA plots for the analysis-group of bead-types comparing single array-sections of 100% UHRR. Although true agreement with intensities from 250 ng is clearly poor when small amounts of starting material are used, the rank-correlation between array-sections remains high even for 10 ng of starting RNA (Table 5), suggesting that it may be possible to normalize samples arising from different starting levels of RNA.

It is also clear from Figure 2 that intensities generally decrease with the quantity of starting RNA, as was observed specifically for the control bead-types. This loss of signal leads to an apparent diminishing of technical biases (e.g. if all signal were lost then we would cease to observe the diminishment of signal as target locations become more 5' along the gene), which should not be mistaken for a benefit.

Differential expression

The number of bead-types identified as showing differential expression ($p < 0.001$, for the non-linear model),

decreases with the amount of starting material much as did the number for which expression was detected (Table 6). Naturally, differential expression implies expression, so we might expect to see the numbers for differential expression bounded by the numbers we saw for expression. The decline in numbers of bead-types for which differential expression is noted is more marked than would be required simply by this constraint. Moreover we should note that due to the nature of the two tests, it is entirely possible to detect differential expression across the set of array-sections, but not detect expression in any individual array-section (Additional File 1, Section 9): evidence that the filtering of bead-types based on expression-detection scores requires caution.

Once more, the sensitivity (defined as for expression detection) is high with a drop-off only when 10 ng of starting RNA are used, and the FDR (defined as for expression detection) remains low across all quantities of starting RNA (Table 7). If we break down the comparison by the magnitude of differential expression (taking 250 ng as the gold standard and comparing the log-expression between 100% UHRR and 100% Brain), then it is apparent (Figure 3) that one pays a price for using the 10 ng level of starting material across the full range of log-fold changes (Additional File 1, Section 10). The performance of the 100 ng and 50 ng starting levels is better, and matches 250 ng outside the range of 0.25 to 1.25. Within that range, they return a lower proportion of bead-types as being differentially expressed, while the 100 ng level of starting material also outperforms the 50 ng level.

Signal to noise

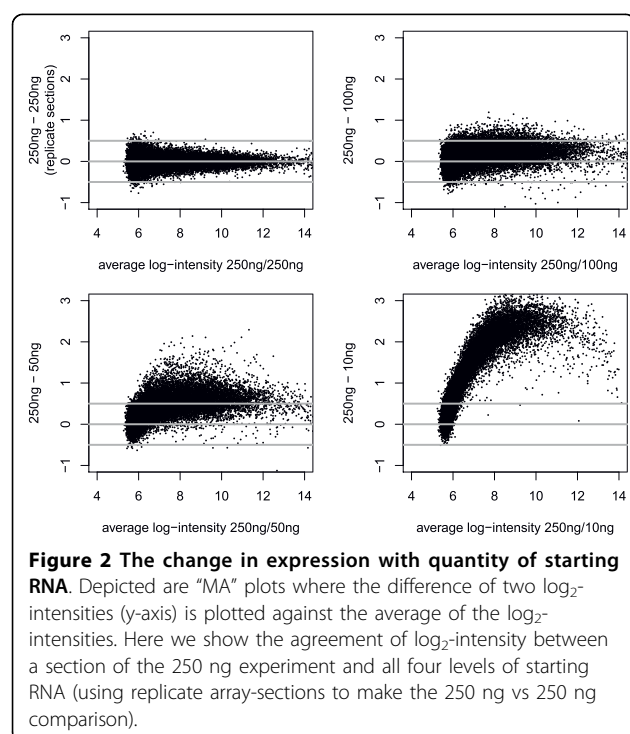
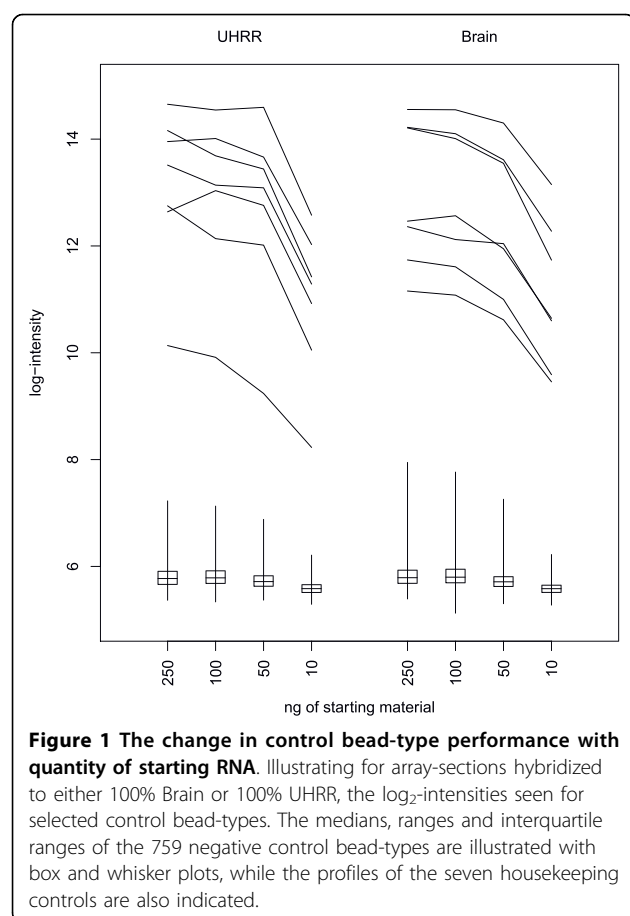
The variance of observations is not independent of their value. Since expression levels decrease as the quantity of starting RNA decreases, it is not possible to assess the change in variance as the quantity of starting RNA decreases, without simultaneously considering the level of expression.

From the non-linear model we can compare the estimate of the difference in expression levels to the estimated

Table 4 Control bead-type summary

Quantity of starting RNA:	250 ng	100 ng	50 ng	10 ng
Negative controls detected in UHRR	37	32	20	0
Negative controls detected in Brain	35	40	29	24
Negative controls detected in all array-sections	10	6	2	0
Negative controls detected in at least one array-section	88	89	98	99
Negative controls: median log-intensity UHRR	5.78	5.79	5.73	5.58
Negative controls: median log-intensity Brain	5.80	5.82	5.72	5.58
Housekeeping controls: median log-intensity UHRR	13.46	13.14	13.09	11.32
Housekeeping controls: median log-intensity Brain	12.47	12.52	12.03	10.64

Giving a) the numbers of the 759 negative-control bead-types that show expression in one or more array-sections, and b) median log-intensities for negative-control and housekeeping bead-types in 100% UHRR and 100% Brain array-sections.



standard error of the difference. This side-steps the complications of the variance and fluorescence levels changing in a dependent manner as the amount of starting material changes. Considering only the analysis-group of bead-types, the median ratios of standard error to estimate are 0.23, 0.28, 0.31 and 0.52 for 250 ng, 100 ng, 50 ng and 10 ng of starting RNA respectively. The median ratios of the two signal to noise ratios are 1.12, 1.16 and 1.76 for 100 ng, 50 ng or 10 ng respectively comparing to a reference starting quantity of 250 ng.

Discussion

Meta-analysis

Inevitably, there will be occasions when we wish to combine data sets generated using different quantities of starting material, possibly because we are performing a meta-analysis across different experiments, or possibly because not all samples within a single experiment can supply the desired quantity of starting RNA. Our analysis has, so far, considered the different quantities of starting material in this study as being different experiments, but we will now briefly consider strategies for combining them.

Consider if samples were run in strata of starting RNA, e.g. we have an experiment where some samples were run using 250 ng, while others were run using 50 ng. The strata were not balanced in terms of experimental design, so we may not wish to obtain two simple estimates for the parameters of interest (one from each stratum) and then combine the estimates. Our strategy for analysis may depend on whether some samples had been run in both strata.

Consider further that we only have Brain run at 50 ng and UHRR at 250 ng, and we wish to transform the 50 ng Brain data for comparison with UHRR. Essentially we wish to simulate a 250 ng Brain data set from this restricted data set, and can use the fact that we do have Brain run at 250 ng to assess the performance. We will consider both the scenario where we have only the two samples with which to work, and a second where we have additionally run UHRR at 50 ng.

If we are in this first scenario, then there is little option but to normalize between the samples. The high rank correlation we have observed between data arising from different starting amounts of RNA gives cause for optimism that a simple quantile-style normalization of the 50 ng data to the expression profile of the 250 ng data will prove successful. With data available from samples run at both starting levels, we can use the 50 ng UHRR and 250 ng UHRR samples to estimate the bias due to starting RNA quantity (via fitting a locally smooth regression) and can then project the 50 ng Brain sample with that model to obtain our prediction for how a 250 ng Brain sample would look. Such an approach shows a marginal improvement over the basic attempt in our example (Figure 4).

Table 5 Squared rank correlations

	Quantity of starting RNA			
	250 ng	100 ng	50 ng	10 ng
250 ng	0.954	0.933	0.921	0.784
100 ng		0.933	0.916	0.789
50 ng			0.924	0.784
10 ng				0.797

Giving the square of Spearman's rank correlation for the intensities of the analysis-group bead-types between 100% UHRR array-sections.

Table 6 Differential expression detected

	250 ng	100 ng	50 ng	10 ng
amongst all bead-types	15,753 (32%)	14,361 (29%)	13,741 (28%)	9,579 (19%)
amongst analysis group	11,021 (51%)	10,169 (47%)	9,788 (45%)	7,084 (33%)

Showing the numbers (and percentage) of bead-types from the complete set of 49,575 for which differential expression between Brain and UHRR was detected. Also showing the same measures for the analysis-group bead-types.

Using the additional data (50 ng UHRR) makes only a small improvement to our ability to transform the 50 ng Brain data and in a real experiment running 50 ng of an additional sample may provide greater value to the ultimate analysis. We should be wary of trying to use a sample for both bias-estimation and analysis as there will be a lack of independence between these samples and all those that are bias-corrected using the results. Moreover the small improvement we see here, over the simpler quantile-normalization style approach, comes using samples that have large numbers of expressed genes. For bias correction of this nature to be useful, we need to observe a wide range of log-intensities which in turn requires large numbers of genes to be expressed. Thus the appropriateness of this more complicated method will be dependent on the size of the experiment and the nature of the samples being hybridized.

Implications for experimental design

A number of implications for experimental design are obvious. It is clear that all things being equal, of the

range of starting quantities of RNA considered here, it is preferable to use 250 ng. If there are limitations to the amount of starting RNA available, then the more starting material used the better (within this range examined). Should the amount of available RNA differ between samples then more subtle decisions are

Table 7 Consistency in detection of differential expression between quantities of starting RNA

test experiment	reference experiment		
	250 ng	100 ng	50 ng
10 ng	0.62/0.04	0.67/0.04	0.69/0.05
50 ng	0.83/0.07	0.88/0.08	
100 ng	0.86/0.06		

For the numbers of bead-types for which differential expression between UHRR and Brain was detected we report "X/Y" where X is the proportion of bead-types reported for the reference experiment also reported for the test experiment (e.g. a measure sensitivity), and Y is the proportion of bead-types reported by the test experiment that were not reported by the reference experiment (e.g. FDR). So for example, taking 250 ng as a gold-standard, for this detection measure the 100 ng experiment has 86% sensitivity and an FDR of 0.06.

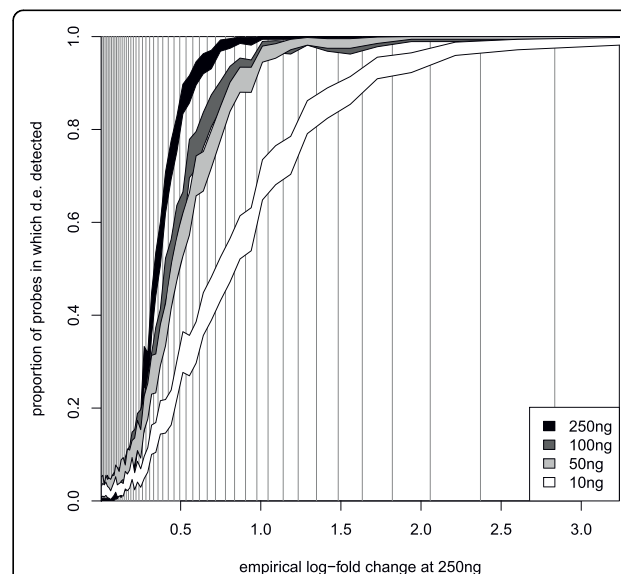


Figure 3 The power to detect differential expression by quantity of starting RNA. Illustrating, for the analysis-group of bead-types, the increased ability to detect large log₂-fold changes (for all levels of starting RNA), and how the relationship (between that ability and the size of the log-fold change) varies with the quantity of starting RNA. The empirical log fold change calculated from the 250 ng experiment is depicted on the x-axis, which is divided into 50 bins, each containing 2% of the bead-types (indicated by the vertical lines). On the y-axis are indicated 95% confidence intervals for the proportion of bead-types in each bin for which differential expression will be detected.

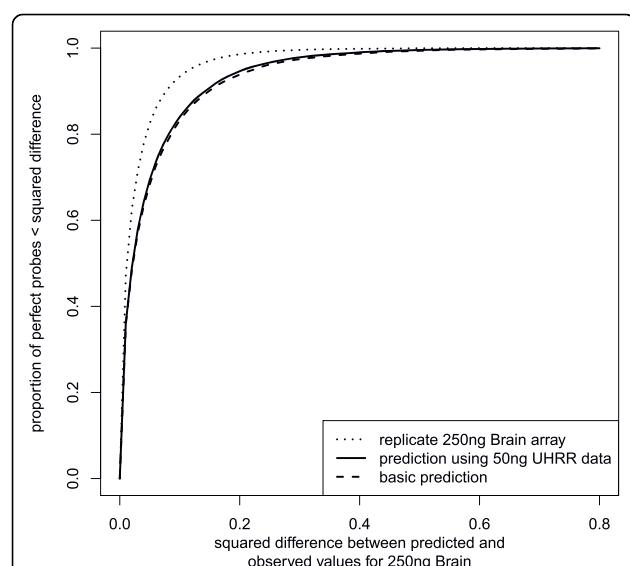


Figure 4 The performance of alternative normalization strategies. The performances of different strategies for normalizing a 50 ng Brain array-section for comparison to a 250 ng UHRR array-section are illustrated. For the analysis-group bead-types, with a 250 ng Brain array-section as a reference, we determine the squared differences between our prediction of \log_2 -intensity and the reference and illustrate the cumulative distribution of those differences. The first prediction uses a basic quantile-style normalization where the 50 ng Brain array-section is transformed to take the distribution of intensities seen on the 250 ng UHRR array-section. A more complicated prediction making use of a 50 ng UHRR section is also illustrated. For reference we give the agreement between replicate 250 ng Brain array-sections, representing the gold-standard that could be achieved by any method.

required. On the basis of the signal-to-noise results, we can infer that if using 100 ng or 50 ng then the sample size would need inflating by a factor of at least 1.2 to achieve the same performance, while if using 10 ng, then in the region of three times the numbers will be required. Thus, when we have the choice and free from other pressures, reducing the starting RNA level is only desirable if it allows sample numbers to be increased by these factors.

The combination of multiple starting RNA levels in one experiment will be problematic. If we wish to normalize using data from the same sample hybridized from multiple quantities of starting RNA, then clearly we must stratify samples into a few starting quantities. If we do not have, or do not wish to make, recourse to replicate samples hybridized from several RNA quantities, and are simply going to normalize samples together, then there is merit in using as much starting RNA as possible for each sample, as was noted in the previous section.

In this scenario, where all samples are independent, it would still be hard to criticize a design that opted for a fixed number of starting levels, especially if this came at

minimal cost to quality (i.e. 250 ng reduced to 220 ng but not to 10 ng) and allowed balance of experimental criteria to be achieved within each stratum of starting quantity. Such an approach is suboptimal by our criterion, but may be more robust to those unexpected events that befall real-world experiments.

Conclusions

We have presented a bead-level Illumina BeadArray dilution control experiment that will be a valuable resource for the Illumina analysis community. As intended, the experiment also answers an important experimental question regarding the required levels of starting RNA, however it also allows for a number of questions to be addressed regarding experimental design when large quantities of RNA are difficult to obtain.

We have shown that reliable signal is obtainable using as little as 10 ng of starting RNA. However we have also seen that lower levels of starting RNA are associated with a bias in expression levels (which may be correctable), and drop in sensitivity (which will not be).

This increase in noise implies that, if using less starting RNA, more samples would be needed in an experiment to achieve the same levels of precision. However, it seems that few false discoveries result from using even as little as 10 ng of starting RNA. Thus while a small experiment using a low starting quantity of RNA may fail to identify many subtle changes, one can have confidence in any changes that are reported.

Additional material

Additional file 1: Supplementary material. File giving details of 1) Experimental Design: Array Layout, 2) Lab Methods: Obtaining bead-level data, 3) Lab Methods: Quality assessment metrics, 4) Lab Methods: Quality assessment - comparison with MAQC, 5) Lab Methods: Quality assessment - Association between starting RNA quantity and intensity, 6) Criteria for including bead-types, 7) Results: Detection, 8) Results: Negative controls, 9) Results: Differential expression but no expression, and 10) Results: Differential expression - detection of small changes.

Acknowledgements

This work was supported by the University of Cambridge, Cancer Research UK [grant number C14303] and Hutchison Whampoa Limited. We thank Nuno Barbosa-Morais for advance access to annotation files. We thank Catherine Ingle, Manolis Dermitzakis and Barbara Stranger of the Wellcome Trust Sanger Institute, Matthew Ritchie of the Walter and Eliza Hall Institute of Medical Research, and Roslin Russell of Cancer Research UK for constructive design discussions.

Author details

¹Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK. ²Cancer Research UK - CRI, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK. ³Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville Victoria 3052, Australia.

Authors' contributions

AGL finalized the design of the experiment, performed the analysis, and drafted the manuscript. JH supervised the experiment and drafted the manuscript. MJD participated in the design of the study, participated in the array data processing and provided Illumina expertise. MO conducted the experiment. NPT and ST participated in the conception and design of the study. All authors read and approved the final manuscript.

Received: 30 April 2010 Accepted: 6 October 2010

Published: 6 October 2010

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**(5235):467-70.
2. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061-8.
3. The ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**(5696):636-40.
4. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurler ME, Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315**(5813):848-53.
5. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays.** *Biostatistics* 2001, **2**(2):183-201.
6. Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nature Reviews Genetics* 2002, **3**(8):579-88.
7. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(25):14863-8.
8. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**(10):R80.
9. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nature Methods* 2009, **6**(5):377-82.
10. UHN Microarray Centre: **Validation of the Illumina iScan System for gene expression.** A UHN microarray centre technical note, University Health Network Microarray Centre, University Health Network Microarray Centre The Toronto Medical Discovery Tower 101 College Street, Rm 9-301 Toronto, Ontario, M5G 1L7 Canada 2009 [http://www.microarrays.ca/info/technical_notes.html].
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics (Oxford, England)* 2003, **4**(2):249-64.
12. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P: **Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms.** *Nucleic Acids Research* 2005, **33**(18):5914-23.
13. Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E, Sun YA, Willey JC, Thierry-Mieg J, Thierry-Mieg D, Setterquist RA, Wilson M, Lucas AB, Novorodovskaya N, Papallo A, Turpaz Y, Baker SC, Warrington JA, Shi L, Herman D: **Using RNA sample titrations to assess microarray platform performance and normalization techniques.** *Nature Biotechnology* 2006, **24**(9):1123-31.
14. Dunning MJ, Barbosa-Morais NL, Tavaré S, Ritchie ME: **Statistical Issues in the analysis of Illumina data.** *BMC Bioinformatics* 2008, **9**(85):1-15.
15. Shi W, Banerjee A, Ritchie ME, Gerondakis S, Smyth GK: **Illumina WG-6 BeadChip strips should be normalized separately.** *BMC Bioinformatics* 2009, **10**:372.
16. Ha KC, Coulombe-Huntington J, Majewski J: **Comparison of Affymetrix Gene Array with the Exon Array shows potential application for detection of transcript isoform variation.** *BMC Genomics* 2009, **10**:519.
17. Mane SP, Evans C, Cooper KL, Crasta OR, Folkerts O, Hutchison SK, Harkins TT, Thierry-Mieg D, Thierry-Mieg J, Jensen RV: **Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing.** *BMC Genomics* 2009, **10**:264.
18. Dunning MJ, Smith ML, Ritchie ME, Tavaré S: **beadarray: R classes and methods for Illumina bead-based data.** *Bioinformatics* 2007, **23**(16):2183-2184.
19. Cairns JM, Dunning MJ, Ritchie ME, Russell R, Lynch AG: **BASH: a tool for managing BeadArray spatial artefacts.** *Bioinformatics* 2008, **24**(24):2921-2922.
20. Holloway AJ, Oshlack A, Diyagama DS, Bowtell DDL, Smyth GK: **Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis.** *BMC Bioinformatics* 2006, **7**:511.
21. Ramdas L, Coombes KR, Baggerly K, Abruzzo L, Highsmith WE, Krogmann T, Hamilton SR, Zhang W: **Sources of nonlinearity in cDNA microarray expression measurements.** *Genome Biology* 2001, **2**(11):11.
22. Zheng X, Huang HC, Li W, Liu P, Li QZ, Liu Y: **Modeling nonlinearity in dilution design microarray data.** *Bioinformatics* 2007, **23**(11):1339-47.
23. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, Tavaré S: **A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data.** *Nucleic Acids Research* 2010, **38**:e17.
24. Smith ML, Dunning MJ, Tavaré S, Lynch AG: **Identification and correction of previously unreported spatial phenomena using raw Illumina BeadArray data.** *BMC Bioinformatics* 2010, **11**:208.

doi:10.1186/1471-2164-11-540

Cite this article as: Lynch *et al.*: The cost of reducing starting RNA quantity for Illumina BeadArrays: A bead-level dilution experiment. *BMC Genomics* 2010 **11**:540.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Chapter 2

Introduction to miRNA Profiling Technologies and Cross-Platform Comparison

Sarah Aldridge and James Hadfield

Abstract

MicroRNA analysis has been widely adopted for basic and applied science. The tools and technologies available for quantifying and analysing miRNAs are still maturing. Here, we give an introductory overview of the main tools and the challenges in their use. We also discuss the importance of basic experimental design, sample handling and analysis methods as the impact of these can be as profound as the choice of miRNA analysis platform. Whether the reader is interested in a gene-by-gene or genome-wide approach choosing the platform to use is not trivial. Careful thought given before starting an experiment will make the execution much easier.

Key words: MicroRNA, Microarray, Sequencing, Reverse transcription quantitative PCR, In situ hybridisation, Comparison

1. Introduction

MicroRNA (miRNA) analysis has rapidly gained a foothold in many labs and is quickly becoming a routine research tool being used in large cohort studies on clinical samples (1), is showing promise in cancer research (2) and has been reported as useful in tumour classification (3) (also reviewed in see ref. 4). MiRNAs hold particular appeal in clinical setting as they have been shown to be very stable in both plasma and serum (5–7). The tools used to measure and detect miRNAs have been largely borrowed from mRNA expression analysis and array-based comparative genomic hybridisation (aCGH) to interrogate DNA copy-number state. The use of microarrays in both mRNA and copy-number is now routine and gradually moving into clinical use; in fact, both techniques are

already being challenged by next-generation sequencing (NGS) methods as the emerging standard method (8, 9). And NGS analysis of miRNA has recently been reported as useful in prediction of clinical outcome (10). Analysis and comparison of different sequencing methods is only just starting to be published (11).

Array-based methods for mRNA and aCGH have been extensively compared (12–14) and the technologies have matured such that they are now routine experimental tools. The different platforms all produce high-quality data and the decisions on choice are often subjectively made. No single platform achieves gold-standard status. Developments in mRNA and aCGH arrays are now primarily an increase in the amount of data generated; i.e., total numbers of probes interrogated. MiRNA analysis techniques have not yet reached the same level of maturity, but the articles in this publication demonstrate how far we have come in a short time and discuss some of the issues yet to be resolved.

2. Problems with miRNA Detection and Quantitation

MiRNAs and other nucleic acids are detected, quantified, and otherwise analysed by three primary methods: hybridisation, PCR, and sequencing. The hybridisation-based Southern (15) and Northern (16) blotting techniques introduced in the 1970s ultimately led to the development of microarrays (17). The polymerase chain reaction (18) was further developed and reverse-transcription quantitative real-time PCR (RT-qPCR) has become the gold-standard technique for nucleic acid quantitation. Sanger sequencing (19) has developed much since its introduction in 1977 and is still the fundamental approach underlying NGS platforms (20–22), even though these systems rely on quite different methods to generate sequence data.

Compared to other nucleic acids however, analysis of miRNA is significantly complicated by several factors: miRNA length, discrimination between pre-, pri-, and mature miRNAs, variable T_m of primers or probes, RNA ligase sequence bias, high degrees of homology in miRNA families and high rates of miRNA discovery. Combinations of these issues impact the different methods for miRNA detection and quantitation and must be considered when designing miRNA experiments. A further complicating factor is that not all the miRNAs present in the central miRNA repository “miRBase” (23) are necessarily real. Resequencing experiments conducted by the Bartel lab (24) found that about 10% of miRBase miRNAs were not present in their dataset and may have been artefactual in other datasets.

A recent review of the major issues with miRNA detection and quantification (25) explains all of the issues listed above and more.

3. miRNA Analysis Technologies

3.1. *Extraction Methods*

The method chosen for preparation of RNAs, including small RNA from cells or tissues is an often overlooked but important aspect of any RNA analysis. The extraction method chosen for RNA, total RNA including the small RNA species or small RNAs alone, can have a downstream effect on the results obtained from a study. Techniques for assessing the quality of small RNA is still an area for which the best practise is still undecided. The impact of methods papers comparing nucleic acid extraction methods is under-valued and this information is often consigned to supplementary methods and commonly never makes it into a formal publication, although there are studies looking into the impact of extraction methodology (26–31). All of these give the very clear and easily understood message that any study should use standardised protocols throughout and employ a single RNA isolation method to avoid the small but potentially significant affects on gene expression analysis due to the RNA preparation method. This is further emphasised in the 2008 review by (25).

Debey et al. and Kim et al. both focussed on the impact of pre-analytical variables, including RNA extraction on gene expression profiling from blood. Debey et al. noted that none of the methods tested outperformed the others. Kim et al. reviewed several studies that tested RNA extraction methods but came to similar conclusions about standardisation. Campo Dell’Orto and Ach et al. both compared three methods of RNA extraction and compared gene expression measurements on arrays and with qRT-PCR. Ach et al. compared: TRIzol (Invitrogen) coupled with isopropanol precipitation, miRNeasy (Qiagen) and mirVana (Life Technologies). They used Agilent miRNA microarrays and real-time PCR to show that very few miRNA gene expression levels were affected by extraction method. Campo Dell’Orto et al. compared: miRNeasy, TRIzol, and TRIzol followed by RNeasy (Qiagen) cleanup. They used Affymetrix HG-U133 Plus 2.0 microarrays to show that the extraction method used does have an impact on gene expression experiments. They suggested the use of a single method but went further in recommending other pre-analytical variables be optimised before gene level analysis. Debey et al. compared: extraction of PBMC cells with TRIzol followed by RNeasy cleanup to whole blood PAXgene (Qiagen) or QIAamp (Qiagen) RNA extraction. They used Affymetrix HGU133A arrays to demonstrate the impact of extraction method on gene expression experiments. Git et al. investigated the best methods for RNA extraction and QC and have used extraction methods based on the Qiagen miRNeasy protocol in their 2010 comparison study. They also performed subsequent yield and quality assessment on Agilent Bioanalyser small RNA series II chips, spectrophotometric analysis and by urea/polyacrylamide gel electrophoresis.

Many of the microarray and sequence-based analysis methods now use total RNA as an input rather than fractionated miRNA. We would very strongly recommend the advice given in all the papers discussed above. A single sample handling and RNA extraction methods should be used in a study. Considering the relatively high cost, and time committed to performing miRNA expression studies efforts should be put into these upstream pre-analytical variables and ideally these should be performed by experienced or practised operators. Users should never compare samples that have been extracted using different methods.

3.2. Reverse Transcription Quantitative PCR

Reverse transcription Quantitative PCR (RT-qPCR) protocols are varied, but essentially rely on conversion of RNA to cDNA and subsequent locus-specific quantification by comparison to a standard reference gene or sample. It is the most sensitive assay technology currently available although re-sequencing may ultimately have equal single copy sensitivity (32). Recently, guidelines for reporting qPCR experiments have been published (33) further strengthening the reliability and intra-lab accessibility of such data.

Methods are available which use either TaqMan probes or SYBR Green. The TaqMan probe-based method (34) starts with a reverse transcription step using gene-specific stem loop primers, which will reverse transcribe both precursor and mature miRNA (35, 36). The alternative SYBR Green-based method (37) uses tagged and anchored oligo- dT primers for reverse transcription of polyadenylated small RNAs for mature miRNAs (38) followed by SYBR Green-based detection. Platforms for medium to high-throughput analysis of miRNA have been an area of intense development of recent years. These have taken the form of assay plates that can assess tens or hundreds of miRNAs across multiple samples in a highly parallel format (39, 40). If large numbers of samples are available for analysis, then RT-qPCR is hard to beat. However, unlike for mRNA, miRNA RT-qPCR is constrained by the detection limitations mentioned above. At least two studies have thus questioned the use of RT-qPCR as a “gold standard” for miRNA quantification (25, 29).

3.3. In Situ Hybridisation

The use of in situ hybridisation (ISH) allows miRNA analysis to be performed directly in tissues of interest and facilitates identification of miRNA expression in specific cell types in complex organs or heterogeneous tumours. Although not a high-throughput tool, ISH can be a very important validation technique once genome-wide miRNA analysis has been conducted. There are several published methods for miRNA ISH, largely using locked nucleic acids (LNA) probes. Probes with LNAs included in the design show increased hybridisation affinities for RNA and miRNA targets over standard probes (41–43). Incorporation of LNAs increases the thermal stability of the probe/RNA complex (44). This is important

as probes for miRNAs need to be short, but the ISH conditions must be stringent to allow for accessibility and hybridisation of relatively short probes (45). Exiqon offers a commercial design service for miRNA ISH probes, which include a proportion of LNAs in the probe sequence. Low signal strength is one downside of LNA probes, but this can be improved by the use of 3' and 5' labelled probes (45).

3.4. Microarrays

Microarrays can be produced in-house (46) or purchased commercially (see Table 1 for a non-exhaustive list). Any array-based method is subject to the same problems of probe design and hybridisation artefacts as described previously. Another problem, discussed in some of the comparison studies, is that not all manufacturers are willing to freely distribute probe sequence information. This data is required for a thorough analysis of miRNA probe characteristics. There is also a risk of obsolescence with microarrays; the Illumina BeadArray and Ambion miRNA platforms were withdrawn in early 2010 both of which performed well in miRNA comparison studies. Users of these products have little control over decisions like this yet comparing results from datasets generated on different platforms is very complex.

The choice of microarray platform is not easy to make. Agilent has almost complete flexibility in array design, Ambion included

Table 1
Platforms analysed in different comparison studies

	RT-qPCR		Microarray platforms										Sequencing	
	SYBR	TaqMan	ABI LDA	Af	Ag	Am	C	E	II	In	L	T	Illumina	SOLiD
Ach		•			•									
Baldwin			•	•	•			•	•				•	•
Chen		•										•		
Dreher		•		•				•		•				
Git	•	•			•	•	•	•	•	•			•	
Pradervand			•	•	•				•				•	
Sah				•	•	•		•	•					
Sato	•				•	•		•		•		•		
Yauk			•		•			•		•	•			

Af Affymetrix, Ag Agilent, Am Ambion, C Combimatrix, E Exiqon, II Illumina, In Invitrogen, L LC sciences, T Toray

putative miRNAs not present in miRBase, Exiqon use LNA to increase specificity, Combimatrix support reuse of arrays, Affymetrix offer a single array containing miRNAs from five species, and other platforms all offer something unique. The technical differences in the available platforms include: printing and surface technology, slide format, labelling, hybridisation, one- or two-colour detection chemistries, probe design, and cost. The input RNA sample requirements also differ widely, from 100 ng of total RNA to 1 µg of small RNA fraction. Replicate spots are useful in downstream analysis and range from 1 to more than 300, with mean spot replicate numbers being from 2 to 5. Surprisingly, the number of replicates is not necessarily constant within a single array platform.

Microarrays are the most obvious choice for users with tens to hundreds of samples for which they wish to perform high-throughput miRNA analysis.

3.5. Next-Generation Sequencing

NGS technology was first used to profile small RNA sequences in *C. elegans* on the 454 platform (47). This study identified several small RNA species and demonstrated that NGS had great potential as a platform for small RNA analysis. Libraries are prepared for NGS using methods based on traditional small RNA cloning techniques. Adapters are ligated to the ends of the small RNA molecules and these are then used as templates for sequencing (48). Small RNA cloning methods for NGS have proved to be technically challenging and time consuming although protocols are improving and alternative methods are becoming available. It is known that biases can be introduced during library production and the implications this has for downstream NGS sequencing has been explored (49, 50). Several steps of the small RNA cloning protocol are noted as hotspots for bias introduction, including adapter ligation, PCR amplification, reverse transcription, and gel isolation.

NGS is particularly well suited to the discovery of novel small RNA species, as the technique is not constrained by the use of hybridisation probes for which prior knowledge of sequence is required. Advances in sequencing technology have accelerated both the discovery rate of new miRNAs and modifications to existing miRNA entries, reflecting subtle variations in mature miRNA sequences (e.g., post-transcriptional editing or terminal residue addition) (51). With the advent of next-generation sequencers with increased capacity for data generation, coupled with advancement in small RNA library preparation methods, many researchers are making use of methods for indexing and multiplexing pools of small RNA libraries to maximise data return. There is a lack of consensus over the best methods for data normalisation, a downside that this platform shares with other methods for small RNA analysis. In addition, associated tools for computational analysis are in their infancy.

4. Microarrays vs. Sequencing

Caveat lector; the discussion below will almost certainly be outdated by the time you read this. We would prefer readers to use this section as a springboard for discussions in their own labs. The rate of change in sequencing technologies is far too great to keep up with in written form. At the time of writing, for instance, Illumina had just announced a 1.14 Tb run on their HiSeq 2000 platform that would allow over 200 exomes in a single run.

There has been much discussion on when, not if, sequencing will supplant microarrays as the analysis method of choice. This discussion is happening almost everywhere that users are running microarrays and is particularly evident on forums like SEQanswers (<http://www.seqanswers.com>). In many cases, the quality of data obtainable from a sequence-based analysis is superior to microarray.

For gene expression analysis, the same levels of detail can be obtained from 10 M sequence reads vs. a standard 3' gene expression array (8, 9). Montgomery et al. showed that this relatively small number of reads produced a similar dynamic range to microarrays but with improved ability to detect and quantify alternatively spliced and very abundant transcripts. Bashir et al. observed that 90% of observed transcripts in a 35 M read dataset can be detected with just 1 M sampled reads, which compares well with the Montgomery et al. analysis. They also noted that an initial sampling run, using highly multiplexed libraries, for instance, could be used as an experimental design tool for transcript sequencing projects. The analysis of alternative splicing has exploded with the advent of NGS. A recent comparison (52) of SOLiD sequencing and Affymetrix exon arrays looked specifically at expression of individual exons, and transcription outside currently annotated loci. They showed that over 80% of exons were detected on both platforms but that RNA-Seq appeared to have a lower background error rate. RNA-Seq was also more sensitive in detecting differentially expressed exons and they could find thousands of novel transcripts with previously unreported exon–exon junctions. Lastly, discovery of new transcripts (mRNAs, miRNAs, LINC RNAs, etc.) is simply not possible using a microarray.

For structural variation analysis, the same levels of detail can also be obtained from about 10 M reads (8). However, as much SV analysis is being done using genotyping intensities from microarrays and the SNP calls bring additional information on LOH that can be used in many studies, there is not a clear choice between the platforms. To obtain the same depth of SNP coverage as an array may require 10–30-fold sequencing of a genome. Sequencing will allow breakpoints and CNV junctions to be mapped to single-nucleotide resolution. Bashir et al. showed that they could resolve

90% of breakpoints using a mix of 200 bp and 2 kb insert size libraries. There is a trade-off between detection and resolution; for a given number of reads increasing library insert size increases the probability of structural variation detection; however, this decreases ultimate resolution of breakpoints. As the number of reads increases in datasets, this issue is reduced. However, many researchers will aim to perform structural variation analysis of tens, possibly hundreds of individuals in a single sequencing run in the near future. An important observation they made was that detection of small structural variations requires the use of libraries with a low insert size distribution and that the distribution must be smaller than the size of the structural variant itself. But even though technologies and methods are improving, long insert library preparation still requires large amounts of nucleic acid. In the case of clinical samples, this can be a major obstacle and the experimental design should balance “sample-cost” vs. structural variation detection and/or resolution.

It is likely that the choice between microarrays and sequencing will be made on secondary factors, such as the platforms locally and easily available.

5. Data Analysis

The use of microarrays for differential miRNA expression led to the adoption of the same or similar tools for their analysis. However, there is an assumption in many mRNA analysis tools that mean mRNA levels are relatively stable and that only a subset of mRNAs might be truly differentially expressed. This is certainly not the case for miRNA analysis, where the number of miRNAs expressed is quite low and there can be stark differences between samples when looking for differentially expressed miRNAs. The methods for processing data can have a similar impact on final results as the technology used in a study (53). Novel methods for miRNA analysis are, and will continue to be developed. Git et al. implemented a novel algorithm to get around the need to choose a reference technology or “gold standard” in their 2010 study.

Understanding the inherent biases in the technique being analysed is important if sensible and biologically meaningful results are to be obtained. The use of spike-in control miRNAs does not necessarily make analysis simpler. However, good experimental design where all variables are considered and a controlled randomised design is used with a single analytical technique will allow useful comparisons to be made. Different biases present in the varied technologies may make certain effects impossible to detect.

6. Comparison Studies

MiRNA analysis methods have only recently been systematically compared (26, 29, 53–59). While microarrays from nine array suppliers were used in these studies none has been used across all those discussed here, one study used six, two used five and four more used three or four array platforms each (Table 1). Since not all studies compared the same arrays, it is somewhat unfair to try and suggest which array platform performs “best”; of course, this is exactly what most readers of these papers want to find out!

Comparisons, and choices, are complicated by the debate about the merits of microarray vs. sequencing vs. real-time PCR as the method of choice. So while there is a large choice of microarray platforms, there is almost as much choice from the non-microarray-based systems (see Table 1). Table 1 shows which platforms were used in the comparison studies we compared. The different platforms generally showed good within- and between-platform reproducibility and correlated well with qPCR, as reported in each study.

Git et al. carried out the most extensive comparison, which encompassed six microarray platforms, real-time PCR using either SYBR Green following reverse transcription with a tagged and anchored oligo-dT primer or TaqMan-based assays with reverse transcription using a pool of gene-specific primers and NGS on the Illumina GAIIX platform.

7. Pitfalls of Comparisons

All comparison studies published have the same flaw; they are outdated as soon as they are available in print. The protocols for sample handling, microarray design or next-gen sequencing technologies improve at a rate far outstripping the ability of authors to produce and analyse comparison datasets. However, these studies are useful to others in deciding which platform to use in a project. Any comparison is likely to reveal shortcomings in the assumptions made about samples, platforms and analysis methods at the start of the process. These may not necessarily be resolvable once the study is complete.

The biological samples chosen and the methods used to extract, quantify, and quality assess them before any biological analysis is made can have a profound effect on the outcome of comparison studies. While many groups have suggested the use of standard samples for use as controls in biological studies, these can only have an impact if these standards are used in the majority of published experiments, which they are not. The samples used in the comparison studies addressed here varied significantly: Ach et al.

used Ambion normal human tissue RNA, HeLa, and ZR-75-1 cell lines; Baldwin et al. used two commercially available RNA samples; Chen et al. used mouse myoblast RNA; Dreher et al. used an HPV-transfected human cell line; Git et al. used an RNA pool from normal breast tissue, and two breast cancer cell lines that were representative of samples used in cancer research; Pradervand et al. used human heart and brain total RNA from Stratagene; Sah et al. used human placenta total RNA spiked with seven synthetic miRNAs in complex pools; Sato et al. used two human RNAs from Ambion; Yauk et al. used two pools of mouse tissue RNA. The majority of these used commercially available RNAs or cell lines that would be relatively easy for others to acquire if they wanted to repeat any aspects of these studies or use them as controls in other work.

A comparison study needs to consider the real-world application of any methods being compared. Protocols for microarray, RT-qPCR, and next-gen sequencing vary from lab to lab. Authors of comparison studies need to decide whether to use manufacturers recommended protocols and starting materials or use their own experience. Both significantly affect the performance of platforms for measuring miRNAs.

The challenges of probe design for miRNAs and the rapidly evolving miRBase database mean that it is important to only compare probes targeting the same miRNA sequence. Several of the comparison papers specifically compared probe sequences and Git et al. commented on the availability, or not, of probe sequence information from the companies compared.

8. Conclusions

Nearly, all technologies used in the comparison studies above performed acceptably in the different measures of performance discussed in each paper. As there are such large differences between and within the available technologies and between platforms, it is important to consider the choice for a particular experiment, and understand that each experimental factor will have an impact on the final results. Comparison studies allow us to quickly assay the performance of a wide variety of systems to measure miRNAs and are of very real benefit to individual scientists. Unfortunately, they do not carry the gravitas of primary scientific publications focussing on biological insights. It would help if these papers were referenced more frequently if the comparison paper aided the choice of platform.

The choice of platform for miRNA analysis needs to balance time, precision, accuracy, cost, and sample type. RT-qPCR is likely to yield the highest sensitivity, use minimal sample and cost least, but is not necessarily practical for profiling hundreds of miRNAs.

Microarrays allow the profiling of tens or hundreds of samples across the known miRNA'ome is shortened from miRNA transcriptome, but are limited by probe design. If it is important to discover new miRNAs, distinguish between isoforms or analyse RNA editing then sequencing is the only method to consider.

Acknowledgements

We thank Stefan Graf, Heidi Dvinge, Claudia Kutter, and Anna Git for their helpful comments on the manuscript.

References

- Volinia, S., Galasso, M., Costinean, S., Tagliavini, L., Gamberoni, G., Drusco, A., et al. (2010). Reprogramming of miRNA networks in cancer and leukemia. *Genome Res* **20**, 589–99.
- Lin, P.-Y., Yu, S.-L., and Yang, P.-C. (2010). MicroRNA in lung cancer. *Br J Cancer* **103**, 1144–8.
- Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* **26**, 462–9.
- Le Quesne, J., and Caldas, C. (2010). MicroRNAs and breast cancer. *Mol Oncol* **4**, 230–41.
- Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci USA* **105**, 10513–18.
- Kroh, E. M., Parkin, R. K., Mitchell, P. S., and Tewari, M. (2010). Analysis of circulating microRNA biomarkers in plasma and serum using quantitative reverse transcription-PCR (qRT-PCR). *Methods* **50**, 298–301.
- Liu, R., Zhang, C., Hu, Z., Li, G., Wang, C., Yang, C., et al. (2010). A five-microRNA signature identified from genome-wide serum microRNA expression profiling serves as a fingerprint for gastric cancer diagnosis. *Eur J Cancer* **47**, 784–91.
- Bashir, A. (2010). Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC Genomics* **11**, 385–99.
- Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–7.
- Hu, Z., Chen, X., Zhao, Y., Tian, T., Jin, G., Shu, Y., et al. (2010). Serum MicroRNA Signatures Identified in a Genome-Wide Serum MicroRNA Expression Profiling Predict Survival of Non-Small-Cell Lung Cancer. *J Clin Oncol* **28**, 1721–6.
- Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709–15.
- Baumbusch, L. O., Aarøe, J., Johansen, F. E., Hicks, J., Sun, H., Bruhn, L., et al. (2008). Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9**, 379–401.
- Curtis, C., Lynch, A. G., Dunning, M. J., Spiteri, I., Marioni, J. C., Hadfield, J., et al. (2009). The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* **10**, 588–601.
- MAQC. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151–61.
- Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* **98**, 503–17.
- Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA* **74**, 5350–4.
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. (1995). Quantitative Monitoring of

- Gene Expression Patterns with a Complementary DNA Microarray *Science* **270**, 467–70.
18. Mullis, K. B. (1990). Target amplification for DNA analysis by the polymerase chain reaction. *Ann Biol Clin (Paris)* **48**, 579–82.
 19. Sanger, F., and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441–8.
 20. Bentley, D. R., Balasubramanian, S., Swerdlow, H., Smith, G. P., Milton, J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–9.
 21. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–80.
 22. McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527–41.
 23. Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154–8.
 24. Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24**, 992–1009.
 25. Nelson, P.T., Wang, W.-X., Wilfred, B. R., and Tang, G. (2009). Technical variables in high-throughput miRNA expression profiling: much work remains to be done. *Biochim Biophys Acta* **1779**, 758–65.
 26. Ach, R. A., Wang, H., and Curry, B. (2008). Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnol* **8**, 69–85.
 27. Campo Dell'Orto, M., Zangrando, A., Trentin, L., Li, R., Liu, W. M., te Kronnie, G., et al. (2007). New data on robustness of gene expression signatures in leukemia: comparison of three distinct total RNA preparation procedures. *BMC Genomics* **8**, 188–203.
 28. Debey, S., Schoenbeck, U., and Hellmich, M. (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *Pharmacogenomics* **4**, 193–207.
 29. Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., et al. (2010). Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16**, 991–1006.
 30. Git, A., Spiteri, I., Blenkiron, C., Dunning, M. J., Pole, J. C., Chin, S. F., et al. (2008). PMC42, a breast progenitor cancer cell line, has normal-like mRNA and microRNA transcriptomes. *Breast Cancer Res* **10**, R54.
 31. Kim, S. J., Dix, D. J., Thompson, K. E., Murrell, R. N., Schmid, J. E., Gallagher, J. E., et al. (2007). Effects of Storage, RNA Extraction, Genechip Type, and Donor Sex on Gene Expression Profiling of Human Whole Blood. *Clin Chem* **53**, 1038–45.
 32. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L., and Quake, S. R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* **105**, 16266–71.
 33. Bustin, S. A., Benes, V., Garson, J. A., Hellems, J., Huggett, J., Kubista, M., et al. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clin Chem* **55**, 611–22.
 34. Heid, C. A., Stevens, J., Livak, K. J., Williams, P. M. (1996). Real time quantitative PCR. *Genome Res* **6**, 986–94.
 35. Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., et al. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* **33**, e179–88.
 36. Schmittgen, T. D., Lee, E. J., Jiang, J., Sarkar, A., Yang, L., Elton, T. S., et al. (2008). Real-time PCR quantification of precursor and mature microRNA. *Methods* **44**, 31–8.
 37. Schneeberger, C., Speiser, P., Kury, F., and Zeillinger, R. (1995). Quantitative detection of reverse transcriptase-PCR products by means of a novel and sensitive DNA stain. *PCR Methods Appl* **4**, 234–8.
 38. Shi, R., and Chiang, V. (2005). Facile means for quantifying microRNA expression by real-time PCR. *Biotechniques* **39**, 519–25.
 39. Morrison, T., Hurley, J., Garcia, J., Yoder, K., Katz, A., Roberts, D., et al. (2006). Nanoliter high throughput quantitative PCR. *Nucleic Acids Res* **34**, e123–31.
 40. Spurgeon, S. L., Jones, R. C., and Ramakrishnan, R. (2008). High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. *PLoS One* **3**, 1662–9.
 41. Kloosterman, W. P., Wienholds, E., Bruijn, E. D., Kauppinen, S., Plasterk, R. H. A. (2006).

- In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat Methods* **3**, 2005–7.
42. Válczi, A., Hornyik, C., Varga, N., Burgyán, J., Kauppinen, S., Havelda, Z. (2004). Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. *Nucleic Acids Res* **32**, e175–82.
 43. Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., et al. (2005). MicroRNA expression in zebrafish embryonic development. *Science* **309**, 310–1.
 44. Thomsen, R., Nielsen, P. S., Jensen, T. H. (2005). Dramatically improved RNA in situ hybridization signals using LNA-modified probes. *RNA* **11**, 1745–8.
 45. Obernosterer, G., Martinez, J., Alenius, M. (2007). Locked nucleic acid-based in situ detection of microRNAs in mouse tissue sections. *Nat Protoc* **2**, 1508–14.
 46. Liu, C. G., Calin, G. A., Meloon, B., Gamliel, N., Sevignani, C., and Ferracin, M., et al. (2004). An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc Natl Acad Sci USA* **101**, 9740–4.
 47. Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., and Nusbaum, C., et al. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–207.
 48. Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., et al. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3–12.
 49. Linsen, S. E., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R. K., et al. (2009). Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**, 474–6.
 50. Tian, G., Yin, X., Luo, H., Xu, X., Bolund, L., and Zhang, X. (2010). Sequencing bias : comparison of different protocols of MicroRNA library construction. *BMC Biotechnol* **10**, 64–73.
 51. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., and Aravin, A., et al. (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129**, 1401–14.
 52. Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., and Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* **11**, 282–302.
 53. Sah, S., McCall, M., Eveleigh, D., Wilson, M., and Irizarry, R. (2010). Performance evaluation of commercial miRNA expression array platforms. *BMC Res Notes* **3**, 80–6.
 54. Baldwin, D. (2009). ABRF microRNA Profiling: Platform Comparison. www.abrf.org/ResearchGroups/Microarray/Activities/R7_Baldwin.pdf.
 55. Chen, Y., Gelfond, J. A., McManus, L. M., and Shireman PK (2009). Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis. *BMC Genomics* **10**, 407–17.
 56. Dreher, A., Rossing, M., Kaczkowski, B., Nielsen, F. C., and Norrild, B. (2010). Differential expression of cellular microRNAs in HPV-11 transfected cells. An analysis by three different array platforms and qRT-PCR. *Biochem Biophys Res Commun* **403**, 357–62.
 57. Pradervand, S., Weber, J., Lemoine, F., Consales, F., Paillusson, A., and Dupasquier, M., et al. (2010). Concordance among digital gene expression, microarrays, and qPCR when measuring differential expression of microRNAs. *Biotechniques* **48**, 219–22.
 58. Sato, F., Tsuchiya, S., Terasawa K., and Tsujimoto G. (2009). Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. *PLoS One* **4**, 5540–52.
 59. Yauk, C., Rowan-Carroll, A., Stead, J., and Williams, A. (2010). Cross-platform analysis of global microRNA expression technologies. *BMC Genomics* **11**, 330–57.

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis^{1,2†*}, Sohrab P. Shah^{3,4*}, Suet-Feung Chin^{1,2*}, Gulisa Turashvili^{3,4*}, Oscar M. Rueda^{1,2}, Mark J. Dunning², Doug Speed^{2,5†}, Andy G. Lynch^{1,2}, Shamith Samarajiwa^{1,2}, Yinyin Yuan^{1,2}, Stefan Gräf^{1,2}, Gavin Ha³, Gholamreza Haffari³, Ali Bashashati³, Roslin Russell², Steven McKinney^{3,4}, METABRIC Group[‡], Anita Langerød⁶, Andrew Green⁷, Elena Provenzano⁸, Gordon Wishart⁸, Sarah Pinder⁹, Peter Watson^{3,4,10}, Florian Markowetz^{1,2}, Leigh Murphy¹⁰, Ian Ellis⁷, Arnie Purushotham^{9,11}, Anne-Lise Børresen-Dale^{6,12}, James D. Brenton^{2,13}, Simon Tavaré^{1,2,5,14}, Carlos Caldas^{1,2,8,13} & Samuel Aparicio^{3,4}

The elucidation of breast cancer subgroups and their molecular drivers requires integrated views of the genome and transcriptome from representative numbers of patients. We present an integrated analysis of copy number and gene expression in a discovery and validation set of 997 and 995 primary breast tumours, respectively, with long-term clinical follow-up. Inherited variants (copy number variants and single nucleotide polymorphisms) and acquired somatic copy number aberrations (CNAs) were associated with expression in ~40% of genes, with the landscape dominated by *cis*- and *trans*-acting CNAs. By delineating expression outlier genes driven in *cis* by CNAs, we identified putative cancer genes, including deletions in *PPP2R2A*, *MTAP* and *MAP2K4*. Unsupervised analysis of paired DNA–RNA profiles revealed novel subgroups with distinct clinical outcomes, which reproduced in the validation cohort. These include a high-risk, oestrogen-receptor-positive 11q13/14 *cis*-acting subgroup and a favourable prognosis subgroup devoid of CNAs. *Trans*-acting aberration hotspots were found to modulate subgroup-specific gene networks, including a TCR deletion-mediated adaptive immune response in the ‘CNA-devoid’ subgroup and a basal-specific chromosome 5 deletion-associated mitotic network. Our results provide a novel molecular stratification of the breast cancer population, derived from the impact of somatic CNAs on the transcriptome.

Inherited genetic variation and acquired genomic aberrations contribute to breast cancer initiation and progression. Although somatically acquired CNAs are the dominant feature of sporadic breast cancers, the driver events that are selected for during tumorigenesis are difficult to elucidate as they co-occur alongside a much larger landscape of random non-pathogenic passenger alterations^{1,2} and germline copy number variants (CNVs). Attempts to define subtypes of breast cancer and to discern possible somatic drivers are still in their relative infancy^{3–6}, in part because breast cancer represents multiple diseases, implying that large numbers (many hundreds or thousands) of patients must be studied. Here we describe an integrated genomic/transcriptomic analysis of breast cancers with long-term clinical outcomes composed of a discovery set of 997 primary tumours and a validation set of 995 tumours from METABRIC (Molecular Taxonomy of Breast Cancer International Consortium).

A breast cancer population genomic resource

We assembled a collection of over 2,000 clinically annotated primary fresh-frozen breast cancer specimens from tumour banks in the UK

and Canada (Supplementary Tables 1–3). Nearly all oestrogen receptor (ER)-positive and/or lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. Additionally, none of the HER2⁺ patients received trastuzumab. As such, the treatments were homogeneous with respect to clinically relevant groupings. An initial set of 997 tumours was analysed as a discovery group and a further set of 995 tumours, for which complete data later became available, was used to test the reproducibility of the integrative clusters (described below). An overview of the main analytical approaches is provided in Supplementary Fig. 1. Details concerning expression and copy number profiling, including sample assignment to the PAM50 intrinsic subtypes^{3,4,7} (Supplementary Fig. 2), copy number analysis (Supplementary Tables 4–8) and validation (Supplementary Figs 3 and 4 and Supplementary Tables 9–11), and *TP53* mutational profiling (Supplementary Fig. 5) are described in the Supplementary Information.

Genome variation affects tumour expression architecture

Genomic variants are considered to act in *cis* when a variant at a locus has an impact on its own expression, or in *trans* when it is associated

¹Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK. ²Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.

³Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. ⁴Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. ⁵Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Cambridge CB3 0WA, UK.

⁶Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Montebello, 0310 Oslo, Norway. ⁷Department of Histopathology, School of Molecular Medical Sciences, University of Nottingham, Nottingham NG5 1PB, UK. ⁸Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. ⁹King's College London, Breakthrough Breast Cancer Research Unit, London WC2R 2LS, UK. ¹⁰Manitoba Institute of Cell Biology, University of Manitoba, Manitoba R3E 0V9, Canada. ¹¹NIHR Comprehensive Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London, London WC2R 2LS, UK. ¹²Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, 0316 Oslo, Norway. ¹³Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK. ¹⁴Molecular and Computational Biology Program, University of Southern California, Los Angeles, California 90089, USA. †Present addresses: Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA (Ch.C.); University College London, Genetics Institute, WC1E 6BT, UK (D.S.).

*These authors contributed equally to this work.

‡Lists of participants and affiliations appear at the end of the paper.

with genes at other sites in the genome. We generated a map of CNAs, CNVs (Supplementary Fig. 6, Supplementary Tables 12–15) and single nucleotide polymorphisms (SNPs) in the breast cancer genome to distinguish germline from somatic variants (see Methods), and to examine the impact of each of these variants on the expression landscape. Previous studies⁸ have shown that most heritable gene expression traits are governed by a combination of *cis* (proximal) loci, defined here as those within a 3-megabase (Mb) window surrounding the gene of interest, and *trans* (distal) loci, defined here as those outside that window. We assessed the relative influence of SNPs, CNVs and CNAs on tumour expression architecture, using each of these variants as a predictor (see Methods) to elucidate expression quantitative trait loci (eQTLs) among patients.

Both germline variants and somatic aberrations were found to influence tumour expression architecture, having an impact on >39% (11,198/28,609) of expression probes genome-wide based on analysis of variance (ANOVA; see Methods), with roughly equal numbers of genes associated in *cis* and *trans*. CNAs were associated with the greatest number of expression profiles (Fig. 1, Supplementary Figs 7–13 and Supplementary Tables 16–20), but were rivalled by SNPs to explain a greater proportion of expression variation on a per-gene basis genome-wide, whereas the contribution from CNVs was more moderate (Fig. 1b and Supplementary Table 21). The true ratio of putative *trans* versus *cis* eQTLs is hard to estimate⁹; however, the large sample size used here allowed the detection of small effects, with 5,401 and 5,462 CNAs significantly (Šidák adjusted *P* value <0.0001) associated in *cis* or in *trans*, respectively. Whereas *cis*-associations tended to be stronger, the *trans*-acting loci modulated a larger number of messenger RNAs, as described below.

Expression outliers refine the breast cancer landscape

As shown above, ~20% of loci exhibit CNA-expression associations in *cis* (Supplementary Fig. 14). To refine this landscape further and identify the putative driver genes, we used profiles of outlying expression (see Methods and ref. 10) and the high resolution and sensitivity of the

Affymetrix SNP 6.0 platform to delineate candidate regions. This approach markedly reduces the complexity of the landscape to 45 regions (frequency > 5, Fig. 2) and narrows the focus, highlighting novel regions that modulate expression. The full enumeration of regions delineated by this approach and their subtype-specific associations (Supplementary Figs 15 and 16 and Supplementary Tables 22–24) includes both known drivers (for example, *ZNF703* (ref. 11), *PTEN* (ref. 12), *MYC*, *CCND1*, *MDM2*, *ERBB2*, *CCNE1* (ref. 13)) and putative driver aberrations (for example, *MDM1*, *MDM4*, *CDK3*, *CDK4*, *CAMK1D*, *PI4KB*, *NCOR1*).

The deletion landscape of breast cancer has been poorly explored, with the exception of *PTEN*. We illustrate three additional regions of significance centred on *PPP2R2A* (8p21, Fig. 2, region 11), *MTAP* (9p21, Fig. 2, region 15) and *MAP2K4* (17p11, Fig. 2, region 33), which exhibit heterozygous and homozygous deletions (Supplementary Figs 15, 17–19 and Supplementary Table 24) that drive expression of these loci. We observe breast cancer subtype-specific (enriched in mitotic ER-positive cancers) loss of transcript expression in *PPP2R2A*, a B-regulatory subunit of the PP2A mitotic exit holoenzyme complex. Somatic mutations in *PPP2R1A* have recently been reported in clear cell ovarian cancers and endometrioid cancers^{14,15}, and methylation silencing of *PPP2R2B* has also been observed in colorectal cancers¹⁶. Thus, dysregulation of specific PPP2R2A functions in luminal B breast cancers adds a significant pathophysiology to this subtype.

MTAP (9p21, a component of methyladenosine salvage) is frequently co-deleted with the *CDKN2A* and *CDKN2B* tumour suppressor genes in a variety of cancers¹⁷ as we observe here (Supplementary Figs 17c and 18). The third deletion encompasses *MAP2K4* (also called *MKK4*) (17p11), a p38/Jun dual specificity serine/threonine protein kinase. *MAP2K4* has been proposed as a recessive cancer gene¹⁸, with mutations noted in cell lines¹⁹. We show, for the first time, the recurrent deletion of *MAP2K4* (Supplementary Figs 17d and 19) concomitant with outlying expression (Supplementary Fig. 15) in predominantly ER-positive cases, and verify homozygous deletions (Supplementary Table 9) in primary tumours, strengthening the evidence for *MAP2K4* as a tumour suppressor in breast cancer.

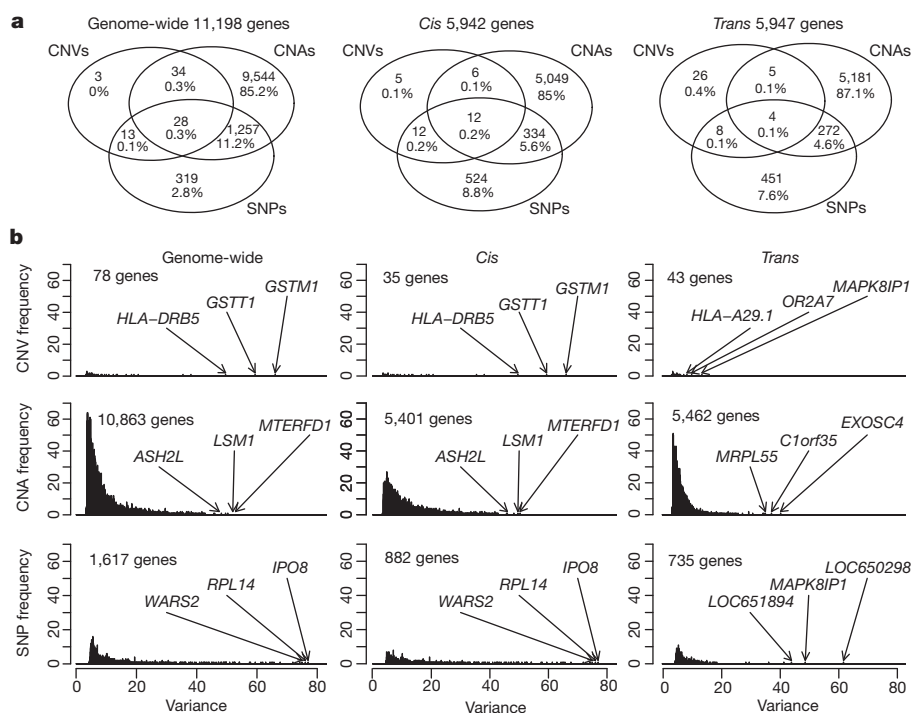


Figure 1 | Germline and somatic variants influence tumour expression architecture. **a**, Venn diagrams depict the relative contribution of SNPs, CNVs and CNAs to genome-wide, *cis* and *trans* tumour expression variation for significant expression associations (Šidák adjusted *P*-value ≤ 0.0001).

b, Histograms illustrate the proportion of variance explained by the most significantly associated predictor for each predictor type, where several of the top associations are indicated.

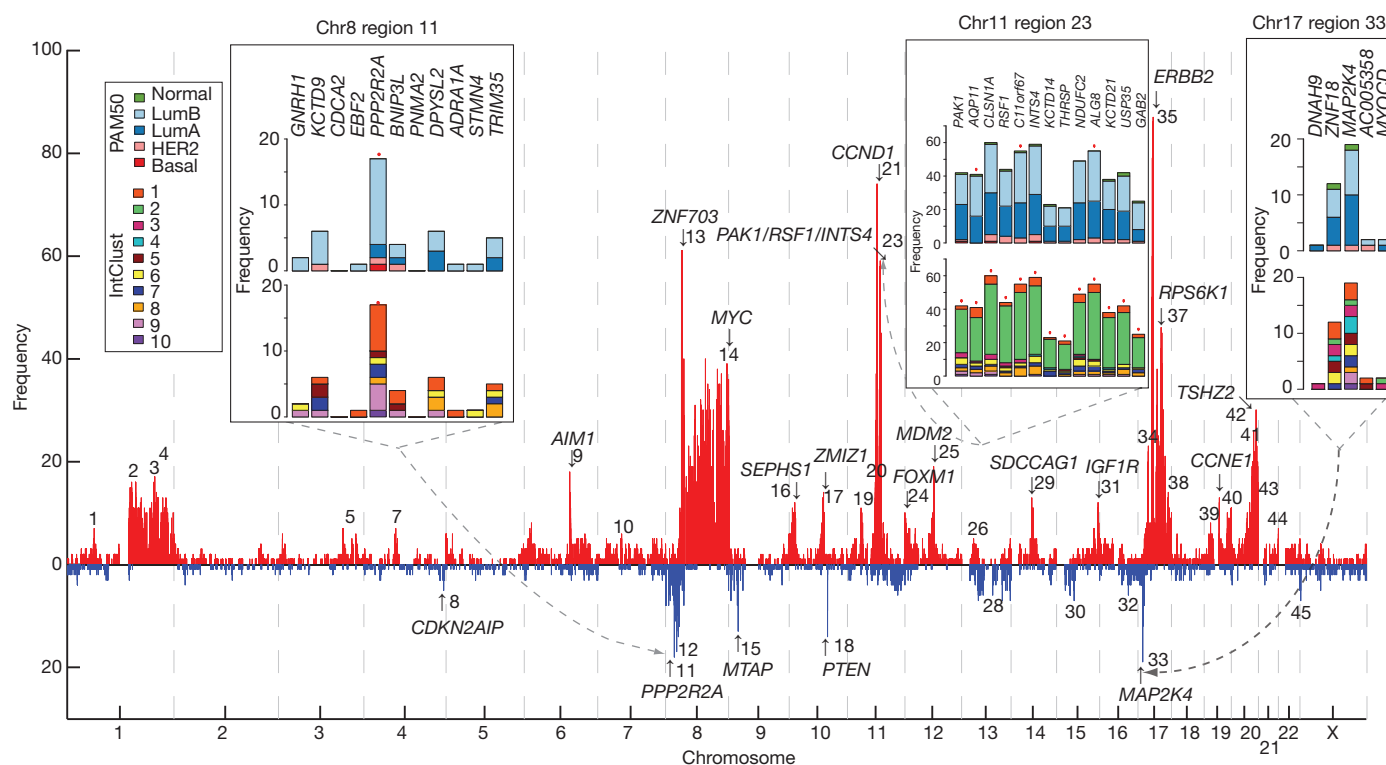


Figure 2 | Patterns of *cis* outlying expression refine putative breast cancer drivers. A genome-wide view of outlying expression coincident with extreme copy number events in the CNA landscape highlights putative driver genes, as indicated by the arrows and numbered regions. The frequency (absolute count) of cases exhibiting an outlying expression profile at regions across the genome is

shown, as is the distribution across subgroups for several regions in the insets. High-level amplifications are indicated in red and homozygous deletions in blue. Red asterisks above the bar plots indicate significantly different observed distributions than expected based on the overall population frequency (χ^2 test, $P < 0.0001$).

Trans-acting associations reveal distinct modules

We next asked how *trans*-associated expression profiles are distributed across the genome. We mapped these in the expression landscape by examining the matrices of CNA–expression associations (see Methods). This revealed strong off-diagonal patterns at loci on chromosomes 1q, 7p, 8, 11q, 14q, 16, 17q and 20q (Fig. 3a), including both positive and negative associations, as well as numerous *trans*-acting aberration hotspots (defined as CNAs associated with >30 mRNAs). Importantly, these aberration hotspots can be grouped into pathway modules, which highlight known driver loci such as *ERBB2* and *MYC*, as well as novel loci associated with large *trans* expression modules (Supplementary Tables 25 and 26). The T-cell-receptor (TCR) loci on chromosomes 7 (*TRG*) and 14 (*TRA*) represent two such hotspots that modulated 381 and 153 unique mRNAs, respectively, as well as 19 dually regulated genes (Supplementary Fig. 20). These cognate mRNAs were highly enriched for T-cell activation and proliferation, dendritic cell presentation, and leukocyte activation, which indicate the induction of an adaptive immune response associated with tumour-infiltrating lymphocytes (Fig. 3b, Supplementary Fig. 20 and Supplementary Tables 27 and 28), as described later.

In a second approach, we examined the genome-wide patterns of linear correlation between copy number and expression features (see Methods), and noted the alignment of several off-diagonal signals, including those on chromosome 1q, 8q, 11q, 14q and 16 (Supplementary Fig. 21). Additionally, a broad signal on chromosome 5 localizing to a deletion event restricted to the basal-like tumours was observed (Supplementary Fig. 21), but was not detected with the eQTL framework, where discrete (as opposed to continuous) copy number values were used. This basal-specific *trans* module is enriched for transcriptional changes involving cell cycle, DNA damage repair and apoptosis (Supplementary Table 29), reflecting the high mitotic index typically associated with basal-like tumours, described in detail below.

Integrative clustering reveals novel subgroups

Using the discovery set of 997 breast cancers, we next asked whether novel biological subgroups could be found by joint clustering of copy number and gene expression data. On the basis of our finding that *cis*-acting CNAs dominated the expression landscape, the top 1,000 *cis*-associated genes across all subtypes (Supplementary Table 30) were used as features for a joint latent variable framework for integrative clustering²⁰ (see Methods). Cluster analysis suggested 10 groups (based on Dunn's index) (see Methods and Supplementary Figs 22 and 23), but for completeness, this result was compared with the results for alternative numbers of clusters and clustering schemes (see Methods, Supplementary Figs 23–27 and Supplementary Tables 31–33). The 10 integrative clusters (labelled IntClust 1–10) were typified by well-defined copy number aberrations (Fig. 4, Supplementary Figs 22, 28–30 and Supplementary Tables 34–39), and split many of the intrinsic subtypes (Supplementary Figs 31–33). Kaplan–Meier plots of disease-specific survival and Cox proportional hazards models indicate subgroups with distinct clinical outcomes (Fig. 5, Supplementary Figs 34, 35 and Supplementary Tables 40 and 41). To validate these results, we trained a classifier (754 features) for the integrative subtypes in the discovery set using the nearest shrunken centroids approach²¹ (see Methods and Supplementary Tables 42 and 43), and then classified the independent validation set of 995 cases into the 10 groups (Supplementary Table 44). The reproducibility of the clusters in the validation set is shown in three ways. First, classification of the validation set resulted in the assignment of a similar proportion of cases to the 10 subgroups, each of which exhibited nearly identical copy number profiles (Fig. 4). Second, the groups have substantially similar hazard ratios (Fig. 5b, Supplementary Fig. 35 and Supplementary Table 40). Third, the quality of the clusters in the validation set is emphasized by the in-group proportions (IGP) measure²² (Fig. 4).

Among the integrative clusters, we first note an ER-positive subgroup composed of 11q13/14 *cis*-acting luminal tumours (IntClust 2,

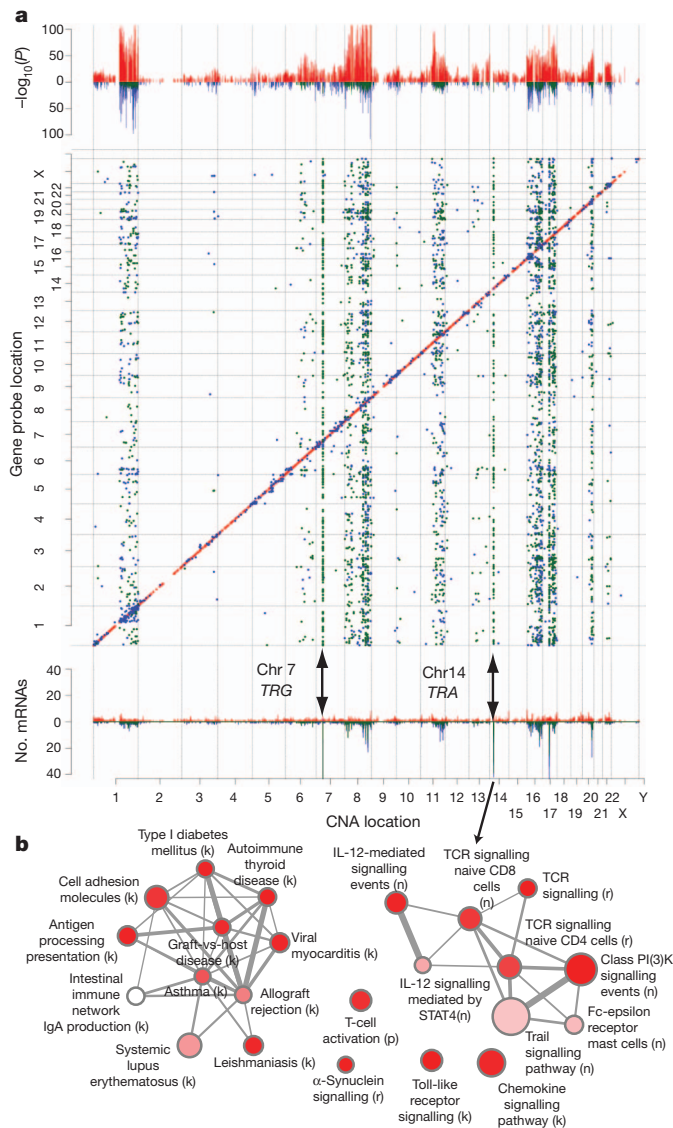


Figure 3 | Trans-acting aberration hotspots modulate concerted molecular pathways. **a**, Manhattan plot illustrating *cis* and *trans* expression-associated copy number aberrations from the eQTL analysis (top panel). The matrix of significant predictor-expression associations (adjusted P -value ≤ 0.0001) exhibits strong off-diagonal patterns (middle panel), and the frequency of mRNAs associated with a particular copy number aberration further illuminates these *trans*-acting aberration hotspots (bottom panel). The directionality of the associations is indicated as follows: *cis*: positive, red; negative, pink; *trans*: positive, blue; negative, green. **b**, Enrichment map of immune response modules in the *trans*-associated TRA network, where letters in parentheses represent the source database as follows: b, NCI-PID BioCarta; c, cancer cell map; k, KEGG; n, NCI-PID curated pathways; p, PANTHER; r, Reactome.

$n = 45$) that harbour other common alterations. This subgroup exhibited a steep mortality trajectory with elevated hazard ratios (discovery set: 3.620, 95% confidence interval (1.905–6.878); validation set: 3.353, 95% confidence interval (1.381–8.141)), indicating that it represents a particularly high-risk subgroup. Several known and putative driver genes reside in this region, namely *CCND1* (11q13.3), *EMSY* (11q13.5), *PAK1* (11q14.1) and *RSF1* (11q14.1), which have been previously linked to breast^{13,23} or ovarian cancer²⁴. Both the copy number (Fig. 4) and expression outlier landscapes (Fig. 2) suggest at least two separate amplicons at 11q13/14, one at *CCND1* (11q13.3) and a separate peak from 11q13.5–11q14.1 spanning *UVRAG*–*GAB2*, centred around *PAK1*, *RSF1*, *C11orf67* and *INTS4*, where it is more challenging to distinguish the driver²⁴. Notably, the

expression outlier profiles for this region are enriched for samples belonging to IntClust 2 (Fig. 2, inset region 23) and all 45 members of this subgroup harboured amplifications of these genes, with high frequencies of amplification also observed for *CCND1* ($n = 39$) and *EMSY* ($n = 34$). In light of these observations, the 11q13/14 amplicon may be driven by a cassette of genes rather than a single oncogene.

Second, we note the existence of two subgroups marked by a paucity of copy number and *cis*-acting alterations. These subgroups cannot be explained by low cellularity tumours (see Methods). One subgroup (IntClust3, $n = 156$) with low genomic instability (Fig. 4 and Supplementary Fig. 22) was composed predominantly of luminal A cases, and was enriched for histotypes that typically have good prognosis, including invasive lobular and tubular carcinomas. The other subgroup (IntClust 4, $n = 167$) was also composed of favourable outcome cases, but included both ER-positive and ER-negative cases and varied intrinsic subtypes, and had an essentially flat copy number landscape, hence termed the ‘CNA-devoid’ subgroup. A significant proportion of cases within this subgroup exhibit extensive lymphocytic infiltration (Supplementary Table 45).

Third, several intermediate prognosis groups of predominantly ER-positive cancers were identified, including a 17q23/20q *cis*-acting luminal B subgroup (IntClust 1, $n = 76$), an 8p12 *cis*-acting luminal subgroup (IntClust 6, $n = 44$), as well as an 8q *cis*-acting/20q-amplified mixed subgroup (IntClust 9, $n = 67$). Two luminal A subgroups with similar CNA profiles and favourable outcome were noted. One subgroup is characterized by the classical 1q gain/16q loss (IntClust 8, $n = 143$), which corresponds to a common translocation event²⁵, and the other lacks the 1q alteration, while maintaining the 16p gain/16q loss with higher frequencies of 8q amplification (IntClust 7, $n = 109$). We also noted that the majority of basal-like tumours formed a stable, mostly high-genomic instability subgroup (IntClust 10, $n = 96$). This subgroup had relatively good long-term outcomes (after 5 years), consistent with ref. 26, and characteristic *cis*-acting alterations (5 loss/8q gain/10p gain/12p gain).

The *ERBB2*-amplified cancers composed of HER2-enriched (ER-negative) cases and luminal (ER-positive) cases appear as IntClust 5 ($n = 94$), thus refining the *ERBB2* intrinsic subtype by grouping additional patients that might benefit from targeted therapy. Patients in this study were enrolled before the general availability of trastuzumab, and as expected this subgroup exhibits the worst disease-specific survival at both 5 and 15 years and elevated hazard ratios (discovery set: 3.899, 95% confidence interval (2.234–6.804); validation set: 4.447, 95% confidence interval (2.284–8.661)).

Pathway deregulation in the integrative subgroups

Finally, we projected the molecular profiles of the integrative subgroups onto pathways to examine possible biological themes among breast cancer subgroups (Supplementary Tables 46 and 47) and the relative impact of *cis* and *trans* expression modules on the pathways. The CNA-devoid (IntClust 4) group exhibits a strong immune and inflammation signature involving the antigen presentation pathway, OX40 signalling, and cytotoxic T-lymphocyte-mediated apoptosis (Supplementary Fig. 36). Given that *trans*-acting deletion hotspots were localized to the *TRG* and *TRA* loci and were associated with an adaptive immune response module, we asked whether these deletions contribute to alterations in this pathway. The CNA-devoid subgroup (IntClust 4) was found to exhibit nearly twice as many deletions (typically heterozygous loss) at the *TRG* and *TRA* loci (~20% of cases) as compared to the other subtypes (with the exception of IntClust 10), and deletions of both TCR loci were significantly associated with severe lymphocytic infiltration (χ^2 test, $P < 10^{-9}$ and $P < 10^{-8}$, respectively). Notably, these *trans*-associated mRNAs were significantly enriched in the immune response signature of the CNA-devoid subgroup (Supplementary Fig. 36) as well as among genes differentially expressed in CNA-devoid cases with severe lymphocytic infiltration (Supplementary Fig. 37). We conclude that genomic copy number loss

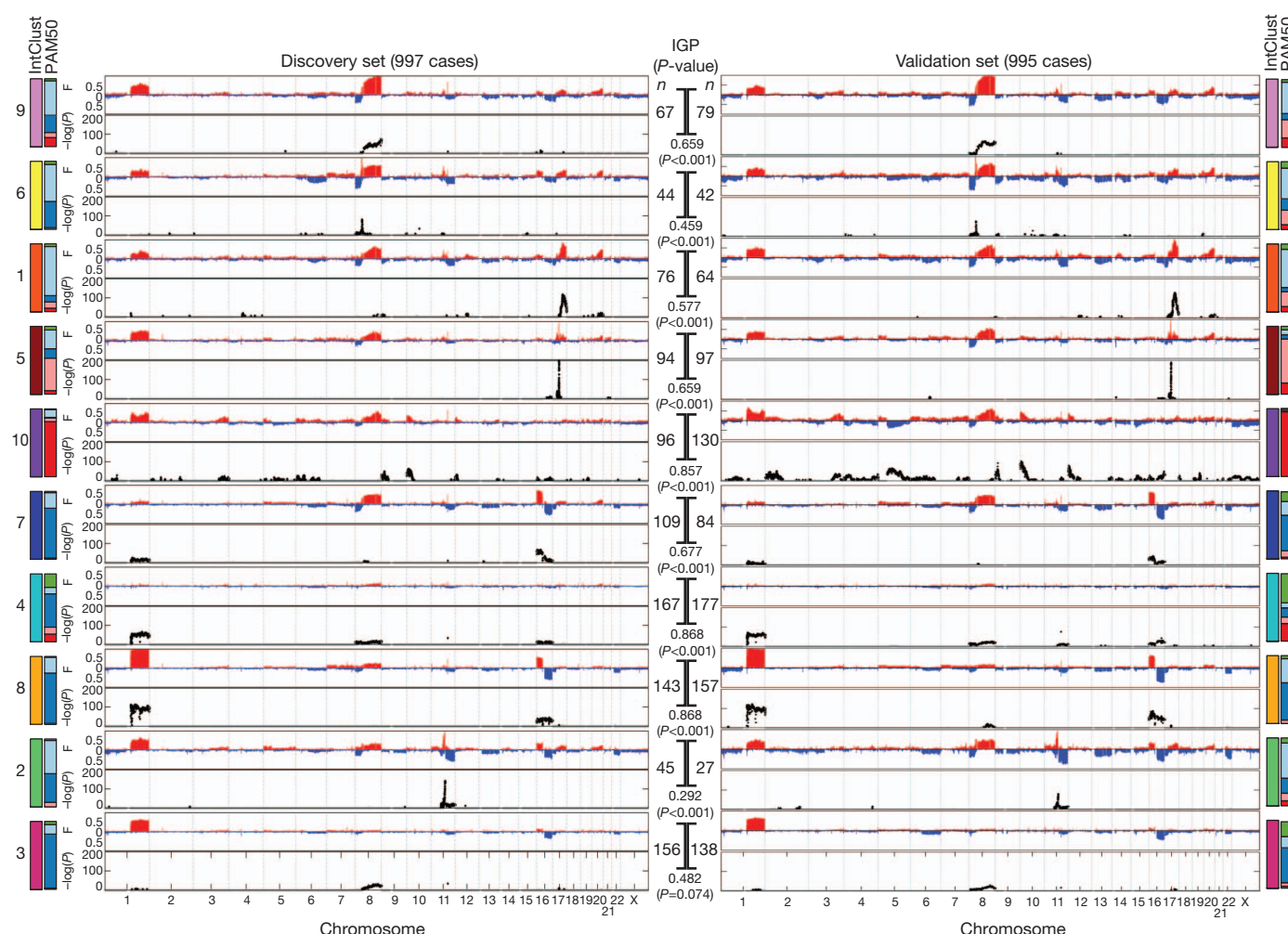


Figure 4 | The integrative subgroups have distinct copy number profiles. Genome-wide frequencies (F, proportion of cases) of somatic CNAs (y-axis, upper plot) and the subtype-specific association ($-\log_{10}$ P-value) of aberrations (y-axis, bottom plot) based on a χ^2 test of independence are shown for each of the 10 integrative clusters. Regions of copy number gain are indicated in red and regions of loss in blue in the frequency plot (upper plot). Subgroups were

ordered by hierarchical clustering of their copy number profiles in the discovery cohort ($n = 997$). For the validation cohort ($n = 995$), samples were classified into each of the integrative clusters as described in the text. The number of cases in each subgroup (n) is indicated as is the in-group proportion (IGP) and associated P-value, as well as the distribution of PAM50 subtypes within each cluster.

at the TCR loci drives a *trans*-acting immune response module that associates with lymphocytic infiltration, and characterizes an otherwise genomically quiescent subgroup of ER-positive and ER-negative patients with good prognosis. These observations suggest the presence of mature T lymphocytes (with rearranged TCR loci), which may explain an immunological response to the cancer. In line with these findings, a recent study²⁷ demonstrated the association between CD8⁺ lymphocytes and favourable prognosis.

Also among the *trans*-influenced groups is IntClust 10 (basal-like cancer enriched subgroup), which harbours chromosome 5q deletions (Supplementary Fig. 21). Numerous signalling molecules, transcription factors and cell division genes were associated in *trans* with this deletion event in the basal cancers, including alterations in *AURKB*, *BCL2*, *BUB1*, *CDCA3*, *CDCA4*, *CDC20*, *CDC45*, *CHEK1*, *FOXM1*, *HDAC2*, *IGF1R*, *KIF2C*, *KIFC1*, *MTHFD1L*, *RAD51API*, *TTK* and *UBE2C* (Supplementary Fig. 38). Notably, *TTK* (*MPS1*), a dual specificity kinase that assists *AURKB* in chromosome alignment during mitosis, and recently reported to promote aneuploidy in breast cancer²⁸, was upregulated. These results indicate that 5q deletions modulate the coordinate transcriptional control of genomic and chromosomal instability and cell cycle regulation within this subgroup.

In contrast to these subtype-specific *trans*-associated signatures, the high-risk 11q13/14 subgroup was characterized by strong

cis-acting associations. Like the basal cancers, this subgroup also exhibited alterations in key cell-cycle-related genes (Supplementary Fig. 39), which probably have a role in its aggressive pathophysiology, but the nature of the signature differs. In particular, the regulation of the G1/S transition by BTG family proteins, which include *CCND1*, *PPP2R1B* and *E2F2*, was significantly enriched in the 11q13/14 *cis*-acting subgroup, but not the basal cancers, and this is consistent with *CCND1* and the *PPP2R* subunit representing subtype-specific drivers in these tumours.

Discussion

We have generated a robust, population-based molecular subgrouping of breast cancer based on multiple genomic views. The size and nature of this cohort made it amenable to eQTL analyses, which can aid the identification of loci that contribute to the disease phenotype²⁹. CNAs and SNPs influenced expression variation, with CNAs dominating the landscape in *cis* and *trans*. The joint clustering of CNAs and gene expression profiles further resolves the considerable heterogeneity of the expression-only subgroups, and highlights a high-risk 11q13/14 *cis*-acting subgroup as well as several other strong *cis*-acting clusters and a genomically quiescent group. The reproducibility of subgroups with these molecular and clinical features in a validation cohort of 995 tumours suggests that by integrating multiple genomic

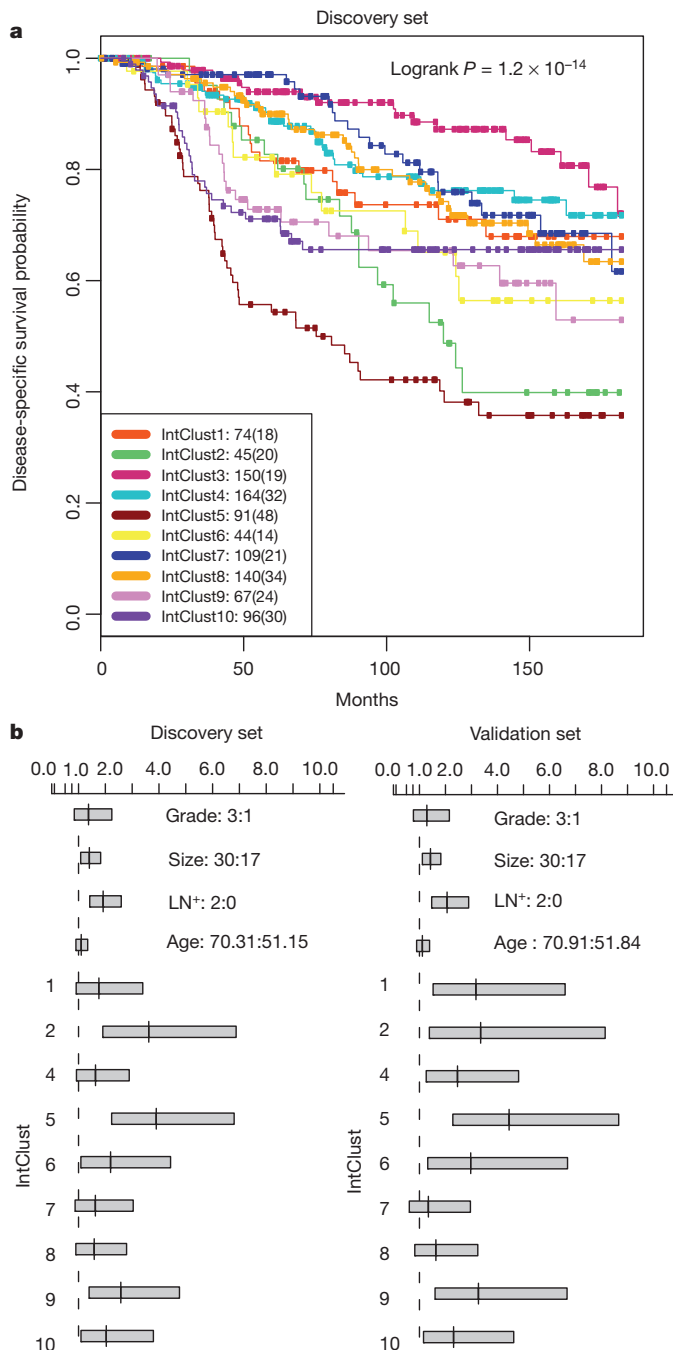


Figure 5 | The integrative subgroups have distinct clinical outcomes.

a, Kaplan-Meier plot of disease-specific survival (truncated at 15 years) for the integrative subgroups in the discovery cohort. For each cluster, the number of samples at risk is indicated as well as the total number of deaths (in parentheses). **b**, 95% confidence intervals for the Cox proportional hazard ratios are illustrated for the discovery and validation cohort for selected values of key covariates, where each subgroup was compared against IntClust 3.

features it may be possible to derive more robust patient classifiers. We show here, for the first time, that subtype-specific *trans*-acting aberrations modulate concerted transcriptional changes, such as the TCR deletion-mediated adaptive immune response that characterizes the CNA-devoid subgroup and the chromosome 5 deletion-associated cell cycle program in the basal cancers.

The integrated CNA-expression landscape highlights a limited number of genomic regions that probably contain driver genes, including *ZNF703*, which we recently described as a luminal B specific driver¹¹, as well as somatic deletion events affecting key subunits of the

PP2A holoenzyme complex and *MTAP*, which have previously been under-explored in breast cancer. The CNA-expression landscape also illuminates rare but potentially significant events, including *IGF1R*, *KRAS* and *EGFR* amplifications and *CDKN2B*, *BRCA2*, *RB1*, *ATM*, *SMAD4*, *NCOR1* and *UTX* homozygous deletions. Although some of these events have low overall frequencies (<1% patients) (Figs 2, Supplementary Fig. 15 and Supplementary Tables 22–24), they may have implications for understanding therapeutic responses to targeted agents, particularly those targeting tyrosine kinases or phosphatases.

Finally, because the integrative subgroups occur at different frequencies in the overall population, focusing sequencing efforts on representative numbers from these groups will help to establish a comprehensive breast cancer somatic landscape at sequence-level resolution. For example, a significant number (~17%, $n = 167$ in the discovery cohort) of breast cancers are devoid of somatic CNAs, and are ripe for mutational profiling. Our work provides a definitive framework for understanding how gene copy number aberrations affect gene expression in breast cancer and reveals novel subgroups that should be the target of future investigation.

METHODS SUMMARY

All patient specimens were obtained with appropriate consent from the relevant institutional review board. DNA and RNA were isolated from samples and hybridized to the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms for genomic and transcriptional profiling, respectively. A detailed description of the experimental assays and analytical methods used to analyse these data are available in the Supplementary Information.

Received 24 April 2011; accepted 22 February 2012.

Published online 18 April 2012.

- Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA* **105**, 16224–16229 (2008).
- Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
- Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Sørli, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
- Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).
- Chin, S. F. *et al.* High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* **8**, R215 (2007).
- Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
- Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
- Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**, 408–415 (2008).
- Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L. & Caldas, C. PACK: Profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics* **22**, 2269–2275 (2006).
- Holland, D. *et al.* *ZNF703* is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol. Med.* **3**, 167–180 (2011).
- Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
- Santarius, T., Shiply, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nature Rev. Cancer* **10**, 59–64 (2010).
- Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
- McConechy, M. K. *et al.* Subtype-specific mutation of *PPP2R1A* in endometrial and ovarian carcinomas. *J. Pathol.* **223**, 567–573 (2011).
- Tan, J. *et al.* B55 β -associated PP2A complex controls PDK1-directed MYC signaling and modulates rapamycin sensitivity in colorectal cancer. *Cancer Cell* **18**, 459–471 (2010).
- Christopher, S. A., Diegelman, P., Porter, C. W. & Kruger, W. D. Methylthioadenosine phosphorylase, a gene frequently codeleted with p16 (CDKN2A/ARF), acts as a tumor suppressor in a breast cancer cell line. *Cancer Res.* **62**, 6639–6644 (2002).
- Teng, D. H. *et al.* Human mitogen-activated protein kinase 4 as a candidate tumor suppressor. *Cancer Res.* **57**, 4177–4182 (1997).
- Hollestelle, A. *et al.* Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Res. Treat.* **121**, 53–64 (2010).

20. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
21. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunk centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
22. Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset? *Biostatistics* **8**, 9–31 (2007).
23. Hughes-Davies, L. *et al.* EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. *Cell* **115**, 523–535 (2003).
24. Brown, L. A. *et al.* Amplification of 11q13 in ovarian carcinoma. *Genes Chromosomes Cancer* **47**, 481–489 (2008).
25. Russnes, H. G. *et al.* Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med.* **2**, 38ra47 (2010).
26. Blows, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med.* **7**, e1000279 (2010).
27. Mahmoud, S. M. A. *et al.* Tumor-infiltrating CD8⁺ lymphocytes predict clinical outcome in breast cancer. *J. Clin. Oncol.* **29**, 1949–1955 (2011).
28. Daniel, J., Coulter, J., Woo, J.-H., Wilsbach, K. & Gabrielson, E. High levels of the Mps1 checkpoint protein are protective of aneuploidy in breast cancer cells. *Proc. Natl Acad. Sci. USA* **108**, 5384–5389 (2011).
29. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The METABRIC project was funded by Cancer Research UK, the British Columbia Cancer Foundation and Canadian Breast Cancer Foundation BC/Yukon. The authors also acknowledge the support of the University of Cambridge, Hutchinson Whampoa, the NIHR Cambridge Biomedical Research Centre, the Cambridge Experimental Cancer Medicine Centre, the Centre for Translational Genomics (CTAG) Vancouver and the BCCA Breast Cancer Outcomes Unit. S.P.S. is a Michael Smith Foundation for Health Research fellow. S.A. is supported by a Canada Research Chair. This work was supported by the National Institutes of Health Centers of Excellence in Genomics Science grant P50 HG02790 (S.T.). The authors thank C. Perou and J. Parker for discussions on the use of the PAM50 centroids. They also acknowledge the patients who donated tissue and the associated pseudo-anonymized clinical data for this project.

Author Contributions Ch.C. led the analysis, designed experiments and wrote the manuscript. S.P.S. led the HMM-based analyses, expression outlier and *TP53* analyses, and contributed to manuscript preparation. S.-F.C. generated data, designed and performed experiments. G.T. generated data, provided histopathology expertise and analysed *TP53* sequence data. O.M.R., M.J.D., D.S., A.G.L., S.S., Y.Y., S.G., Ga.H., Gh.H., A.B., R.R., S.M. and F.M. performed analyses. G.T., A.G., E.P., S.P. and I.E. provided histopathology expertise. A.L. performed *TP53* sequencing. A.-L.B.-D. oversaw *TP53* sequencing. S.P., P.W., L.M., G.W., I.E., A.P., Ca.C. and S.A. contributed to sample selection. J.D.B. and S.T. contributed to study design. S.T. provided statistical expertise. The METABRIC Group contributed collectively to this study. Ca.C. and S.A. co-conceived and oversaw the study, and contributed to manuscript preparation and were responsible for final editing. Ca.C. and S.A. are joint senior authors and project co-leaders.

Author Information The associated genotype and expression data have been deposited at the European Genome-Phenome Archive (<http://www.ebi.ac.uk/ega/>), which is hosted by the European Bioinformatics Institute, under accession number EGAS00000000083. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Ca.C. (carlos.caldas@cancer.org.uk) or S.A. (saparicio@bccrc.ca).

METABRIC Group

Co-chairs Carlos Caldas^{1,2}, Samuel Aparicio^{3,4}

Writing committee Christina Curtis^{1,2†}, Sohrab P. Shah^{3,4}, Carlos Caldas^{1,2}, Samuel Aparicio^{3,4}

Steering committee James D. Brenton^{1,2}, Ian Ellis⁵, David Huntsman^{3,4}, Sarah Pinder⁶, Arnie Purushotham⁶, Leigh Murphy⁷, Carlos Caldas^{1,2}, Samuel Aparicio^{3,4}

Tissue and clinical data source sites: University of Cambridge/Cancer Research UK Cambridge Research Institute Carlos Caldas (Principal Investigator)^{1,2}; Helen Bardwell², Suet-Feung Chin^{1,2}, Christina Curtis^{1,2†}, Zhihao Ding², Stefan Gräf^{1,2}, Linda Jones⁸, Bin Liu^{1,2}, Andy G. Lynch^{1,2}, Irene Papatheodorou^{1,2}, Stephen J. Sammut⁹, Gordon Wishart⁹, **British Columbia Cancer Agency** Samuel Aparicio (Principal Investigator)^{3,4}, Steven Chia⁴, Karen Gelmon⁴, David Huntsman^{3,4}, Steven McKinney^{3,4}, Caroline Speers⁴, Gulisa Turashvili^{3,4}, Peter Watson^{3,4,7}; **University of Nottingham**: Ian Ellis (Principal Investigator)⁵, Roger Blamey⁵, Andrew Green⁵, Douglas Macmillan⁵, Emad Rakha⁵; **King's College London** Arnie Purushotham (Principal Investigator)⁶, Cheryl Gillett⁶, Anita Grigoriadis⁶, Sarah Pinder⁶, Emanuele di Rinaldis⁶, Andy Tutt⁶; **Manitoba Institute of Cell Biology** Leigh Murphy (Principal Investigator)⁷, Michelle Parisien⁷, Sandra Troup⁷

Cancer genome/transcriptome characterization centres: University of Cambridge/Cancer Research UK Cambridge Research Institute Carlos Caldas (Principal Investigator)^{1,2}, Suet-Feung Chin (Team Leader)^{1,2}, Derek Chan¹, Claire Fielding², Ana-Teresa Maia^{1,2}, Sarah McGuire², Michelle Osborne², Sara M. Sayalero², Inmaculada Spiteri², James Hadfield²; **British Columbia Cancer Agency** Samuel Aparicio (Principal Investigator)^{3,4}, Gulisa Turashvili (Team Leader)^{3,4}, Lynda Bell⁴, Katie Chow⁴, Nadia Gale⁴, David Huntsman^{3,4}, Maria Kovalik⁴, Ying Ng⁴, Leah Prentice⁴

Data analysis subgroup: University of Cambridge/Cancer Research UK Cambridge Research Institute Carlos Caldas (Principal Investigator)^{1,2}, Simon Tavaré (Principal Investigator)^{1,2,10,11}, Christina Curtis (Team Leader)^{1,2†}, Mark J. Dunning², Stefan Gräf^{1,2}, Andy G. Lynch^{1,2}, Oscar M. Rueda^{1,2}, Roslin Russell², Shamith Samarajiva^{1,2}, Doug Speed^{2,10}, Florian Markowetz (Principal Investigator)^{1,2}, Yinyin Yuan^{1,2}; James D. Brenton (Principal Investigator)^{1,2}; **British Columbia Cancer Agency** Samuel Aparicio (Principal Investigator)^{3,4}, Sohrab P. Shah (Team Leader)^{3,4}, Ali Bashashati³, Gavin Ha³, Gholamreza Haffari³ & Steven McKinney^{3,4}

¹Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK. ²Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ³Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. ⁴Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. ⁵Department of Histopathology, School of Molecular Medical Sciences, University of Nottingham, Nottingham NG5 1PB, UK. ⁶King's College London, Breakthrough Breast Cancer Research Unit, London, WC2R 2LS, UK. ⁷Manitoba Institute of Cell Biology, University of Manitoba, Manitoba R3E 0V9, Canada. ⁸Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK. ⁹Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. ¹⁰Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Cambridge CB3 0WA, UK. ¹¹Molecular and Computational Biology Program, University of Southern California, Los Angeles, California 90089, USA. †Present address: Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA.

CANCER GENOMICS

Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA

Tim Forsheew,^{1*} Muhammed Murtaza,^{1,2*} Christine Parkinson,^{1,2,3*} Davina Gale,^{1*} Dana W. Y. Tsui,^{1*} Fiona Kaper,^{4†} Sarah-Jane Dawson,^{1,2,3} Anna M. Piskorz,^{1,2} Mercedes Jimenez-Linan,^{3,5} David Bentley,⁶ James Hadfield,¹ Andrew P. May,⁴ Carlos Caldas,^{1,2,3,7} James D. Brenton,^{1,2,3,7‡} Nitzan Rosenfeld^{1,2‡}

Plasma of cancer patients contains cell-free tumor DNA that carries information on tumor mutations and tumor burden. Individual mutations have been probed using allele-specific assays, but sequencing of entire genes to detect cancer mutations in circulating DNA has not been demonstrated. We developed a method for tagged-amplicon deep sequencing (TAm-Seq) and screened 5995 genomic bases for low-frequency mutations. Using this method, we identified cancer mutations present in circulating DNA at allele frequencies as low as 2%, with sensitivity and specificity of >97%. We identified mutations throughout the tumor suppressor gene *TP53* in circulating DNA from 46 plasma samples of advanced ovarian cancer patients. We demonstrated use of TAm-Seq to noninvasively identify the origin of metastatic relapse in a patient with multiple primary tumors. In another case, we identified in plasma an *EGFR* mutation not found in an initial ovarian biopsy. We further used TAm-Seq to monitor tumor dynamics, and tracked 10 concomitant mutations in plasma of a metastatic breast cancer patient over 16 months. This low-cost, high-throughput method could facilitate analysis of circulating DNA as a noninvasive “liquid biopsy” for personalized cancer genomics.

INTRODUCTION

Circulating cell-free DNA extracted from plasma or other body fluids has potentially transformative applications in cancer management (1–7). Characterization of tumor mutation profiles is required for informed choice of therapy, given that biological agents target specific pathways and effectiveness may be modulated by specific mutations (8–11). Yet, mutation profiles in different metastatic clones can differ significantly from each other or from the parent primary tumor (12, 13). Evolutionary changes within the cancer can alter the mutational spectrum of the disease and its responsiveness to therapies, which may necessitate repeat biopsies (14–17). Biopsies are invasive and costly and only provide a snapshot of mutations present at a given time and location. For some applications, mutation detection in plasma DNA as a “liquid biopsy” could potentially replace invasive biopsies as a means to assess tumor genetic characteristics (2–7). Sensitive methods for detecting cancer mutations in plasma may find use in early detection screening (1), prognosis, monitoring tumor dynamics over time, or detection of minimal residual disease (3, 18, 19). In high-grade serous

ovarian carcinomas (HGSOC), mutations in the tumor suppressor gene *TP53* have been observed in 97% of cases (20, 21), but these are located throughout the gene and are difficult to assay. A cost-effective method that could detect and measure allele frequency (AF) of *TP53* mutations in plasma may be highly applicable as a biomarker for HGSOC (22).

Circulating DNA is fragmented to an average length of 140 to 170 base pairs (bp) and is present in only a few thousand amplifiable copies per milliliter of blood, of which only a fraction may be diagnostically relevant (2, 3, 23–25). Recent advances in noninvasive prenatal diagnostics highlight the clinical potential of circulating DNA (25–28), but also the challenges involved in analysis of circulating tumor DNA (ctDNA), where mutated loci and AFs may be more variable. Various methods have been optimized to detect extremely rare alleles (1, 2, 6, 7, 29–31), and can assay for predefined or hotspot mutations. These methods, however, interrogate individual or few loci and have limited ability to identify mutations in genes that lack mutation hotspots, such as the *TP53* and *PTEN* tumor suppressor genes (32). In patients with more advanced cancers, ctDNA can comprise as much as 1% to 10% or more of circulating DNA (2), presenting an opportunity for more extensive genomic analysis. Targeted resequencing has been recently used to identify mutations in selected genes at AFs as low as 5% (33–35). However, identifying mutations across sizeable genomic regions spanning entire genes at an AF as low as 2%, or in few nanograms of fragmented template from circulating DNA, has been more challenging.

In response, we describe a tool for noninvasive mutation analysis on the basis of tagged-amplicon deep sequencing (TAm-Seq), which allows amplification and deep sequencing of genomic regions spanning thousands of bases from as little as individual copies of fragmented DNA. We applied this technique for detection of both abundant and

¹Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ²Department of Oncology, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK. ³Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. ⁴Fluidigm Corporation, 7000 Shoreline Court, Suite 100, South San Francisco, CA 94080, USA. ⁵Department of Histopathology, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. ⁶Illumina Cambridge, Chesterford Research Park, Little Chesterford, Cambridge CB10 1XL, UK. ⁷Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.

*These authors contributed equally to this work.

†Present address: Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA.

‡To whom correspondence should be addressed. E-mail: nitzan.rosenfeld@cancer.org.uk (N.R.); james.brenton@cancer.org.uk (J.D.B.)

rare mutations in circulating DNA from blood plasma of ovarian and breast cancer patients. This sequencing approach allowed us to monitor changes in tumor burden by sampling only patient plasma over time. Combined with faster, more accurate sequencing technologies or rare allele amplification strategies, this approach could potentially be used for personalized medicine at point of care.

RESULTS

Targeted deep sequencing of fragmented DNA by TAM-Seq

To amplify and sequence fragmented DNA, we designed primers to generate amplicons that tile regions of interest in short segments of about 150 to 200 bases (Fig. 1A and table S1), incorporating universal

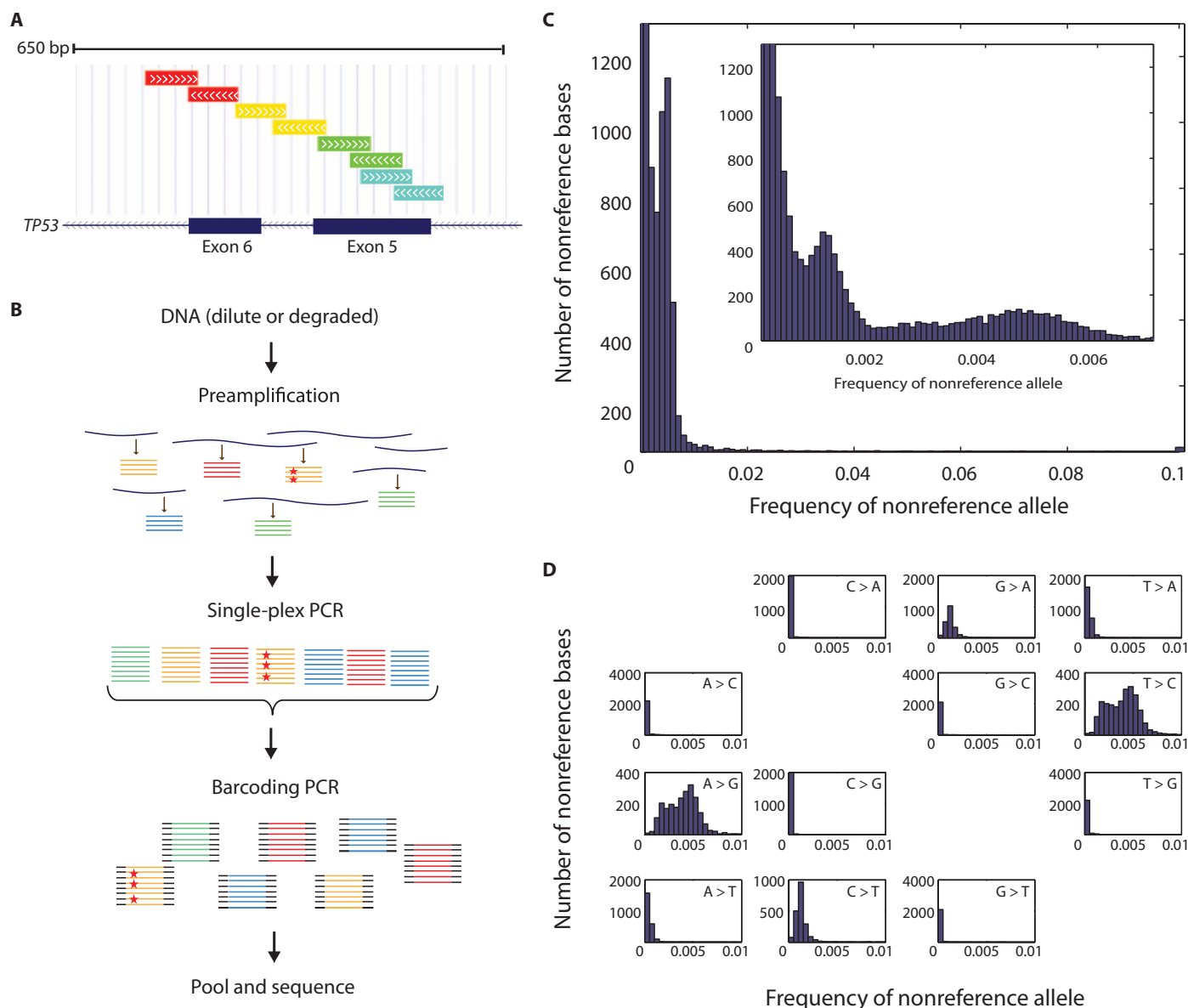


Fig. 1. Overview of tagged amplicon sequencing (TAM-Seq). **(A)** Illustration of amplicon design. Primers were designed to amplify regions of interest in overlapping short amplicons (table S1). Amplicon design is illustrated for a region covering exons 5 to 6 of *TP53*. Colored bars, segmented into forward and reverse reads, show regions covered by different amplicons (excluding primer regions). Sequencing adaptors are attached at either end, such that a single-end read generates separate sets of forward and reverse reads (fig. S1). Because amplicons are mostly shorter than 200 bp, the forward and reverse reads also partially overlap. Figure adapted from University of California, Santa Cruz, Genome Browser (<http://genome.ucsc.edu/>). **(B)** Workflow overview. Multiple regions were amplified in parallel. An initial preamplification step was

performed for 15 cycles using a pool of the target-specific primer pairs to preserve representation of all alleles in the template material. The schematic diagram shows DNA molecules that carry mutations (red stars) being amplified alongside wild-type molecules. Regions of interest in the preamplified material were then selectively amplified in individual (single-plex) PCR, thus excluding nonspecific products. Finally, sequencing adaptors and sample-specific barcodes were attached to the harvested amplicons in a further PCR. **(C)** Distribution of observed nonreference read frequencies, averaged over 47 FFPE samples, across all loci and all nonreference bases. Inset expands the low-frequency range. **(D)** Distribution of the observed background nonreference read frequencies averaged over 47 FFPE samples for the 12 different A/C/G/T base substitutions.

adaptors at 5' ends (fig. S1). Performing single-plex amplification with each of these primer pairs would require dispersing the initial sample into many separate reactions, considerably increasing the probability of sampling errors and allelic loss. Multiplex amplification using a large set of primers could result in nonspecific amplification products and biased coverage. We therefore applied a two-step amplification process: a limited-cycle preamplification step where all primer sets were used together to capture the starting molecules present in the template, followed by individual amplification to purify and select for intended targets (Fig. 1B) (Supplementary Methods). The final concentration of each primer in the preamplification reaction was 50 nM, reducing the potential for interprimer interactions, and 15 cycles of long-extension (4 min) polymerase chain reaction (PCR) were used to remain in the exponential phase of amplification. We used a microfluidic system (Access Array, Fluidigm) to perform parallel single-plex amplification from multiple preamplified samples using multiple primer sets. An additional PCR step attached sequencing adaptors (fig. S1) and tagged each sample by a unique molecular identifier or "barcode" (table S2). Sequencing adaptors were separately attached at either end and the products mixed together, such that single-end sequencing generated separate sets of forward and reverse reads. We performed 100-base single-end sequencing (GAIIx sequencer, Illumina), with an additional 10 cycles using the barcode sequencing primer, generating ~30 million reads per lane. This produced an average read depth of 3250 for each of 96 barcoded samples for 48 amplicons read in two possible orientations.

Validation and sensitivity for mutation identification in ovarian tumor samples

We designed a set of 48 primer pairs to amplify 5995 bases of genomic sequence covering coding regions (exons and exon junctions) of *TP53* and *PTEN*, and selected regions in *EGFR*, *BRAF*, *KRAS*, and *PIK3CA* (table S1) by overlapping short amplicons (Fig. 1A). The sequenced regions cover mutations that account for 38% of all point mutations in the COSMIC database (v55) (32). We used TAM-Seq to sequence DNA extracted from 47 formalin-fixed, paraffin-embedded (FFPE) tumor specimens of ovarian cancers (table S3), which were also sequenced for *TP53* by Sanger sequencing (36) (Supplementary Methods). DNA extracted from FFPE samples is generally degraded and fragmented as a result of fixation and long-term ambient storage. We amplified DNA from each sample in duplicate, tagging each replicate with a different barcode. Using a single lane of sequencing, we generated 3.5 gigabases of data passing signal purity filters, producing mean read depth of 3200 above Q30 for each of the 9024 expected read groups (48 amplicons \times 2 directions \times 94 barcoded samples). Background frequencies of nonreference reads were ~0.1% (median, 0.03%; mean, 0.2%; in keeping with Q30 quality threshold applied), yet varied substantially between loci and base substitutions (Fig. 1C) and showed a clear bias toward purine/pyrimidine conservation (Fig. 1D). Sixty-six percent of loci had mean background rate of <0.1%, and 96% of loci had background rate of <0.6%.

The data set interrogated nearly 18,000 possible single-base substitutions for each sample, which introduces a risk of false detection. To control for sporadic PCR errors and reduce false positives, we called point mutations in a sample only if nonreference AFs were above the respective substitution-specific background distribution at a high confidence margin (0.9995 or greater), and ranked high in the list of nonreference AFs, in both replicates (Supplementary Methods). Duplicate

sequencing data were obtained for 44 samples, and 43 single-base substitutions were called (table S3). These matched 100% of mutations identified by Sanger sequencing and included three additional mutations at low AFs that were below detection thresholds of Sanger sequencing (fig. S2). The upper bound of AFs that may have been missed was estimated (Supplementary Methods) at <5% for 36 of 44 FFPE samples (82%) and <10% for 42 of 44 samples (95%), with median value of 1.3% and mean value of 2.7%. Mutant AFs were highly reproducible in duplicate samples. For 42 of 43 mutations called, the difference in measured frequency between duplicates was less than 0.08, and the relative difference was 25% or less (Fig. 2A). Mutant AFs correlated significantly with tumor cellularity in the FFPE block (correlation coefficient = 0.422; $P = 0.0049$, t test) (Fig. 2B).

In a separate run, we sequenced libraries prepared from six different diluted mixtures of six FFPE samples, with a different known point mutation in *TP53* in each, to mean read depth of 5600. Of more than 100,000 possible non-SNP (single-nucleotide polymorphism) substitutions, we identified all 33 expected point mutations present at AF >1%, including 6 mutations present at AF <2%, with one false-positive called with AF = 1.9%. Using less stringent parameters (Supplementary Methods), we identified three additional mutations present at AF = 0.6% (Fig. 2C), with no additional false positives. Thus, we obtained 100% sensitivity, identifying mutations at AFs as low as 0.6%. A positive predictive value (PPV) of 100% was calculated for mutations at AF >2%, and a PPV of 90% for mutations identified at AF <2% (Fig. 2D).

Quantitative limitations of mutation detection

When applying TAM-Seq to measure a predefined mutation (as opposed to screening thousands of possible substitutions), the frequency of the mutant allele can be read out directly from the data at the desired locus. False detection is less likely, and criteria for confident mutation detection for a predefined substitution can be less stringent than those described above for de novo mutation identification (Supplementary Methods). The minimal nonreference AFs that could be detected depend on the read depth and background rates of nonreference reads, which vary per locus and substitution type. Minimal detectable frequencies increase when higher confidence margins are used (Supplementary Methods) and had a median value of 0.14% at confidence margin of 0.95 and 0.18% at confidence margin of 0.99 (fig. S3). The minimal detectable frequency would also be limited if a minimal number of reads is applied for confident mutation detection; for example, a minimum of 10 reads implies that sequencing depth of 5000 would be required to detect mutations at AF as low as 0.2%. For alleles present at ~10 or fewer copies in the starting template, reproducibility would also be limited by sampling noise, because these alleles may be over- or underrepresented in any particular reaction.

To characterize the quantitative accuracy of TAM-Seq as applied to circulating DNA, we simulated rare circulating tumor mutations by mixing plasma DNA from two healthy individuals. Using the same set of primers as used for the FFPE experiment, we identified that these two individuals differed at five known SNP loci (table S4). Total amplifiable copies in both plasma DNA samples were determined by digital PCR and mixed to obtain minor AFs ranging from 0.16% to 40% (Supplementary Methods). We sequenced diluted templates containing between 250 and <1 expected copy of the minor allele (table S5). The coefficient of variation (CV) of the observed AFs was equal on average to the inverse square root ($1/\sqrt{n}$) of the expected number of copies of the rare allele (Fig. 3A), which is the theoretical

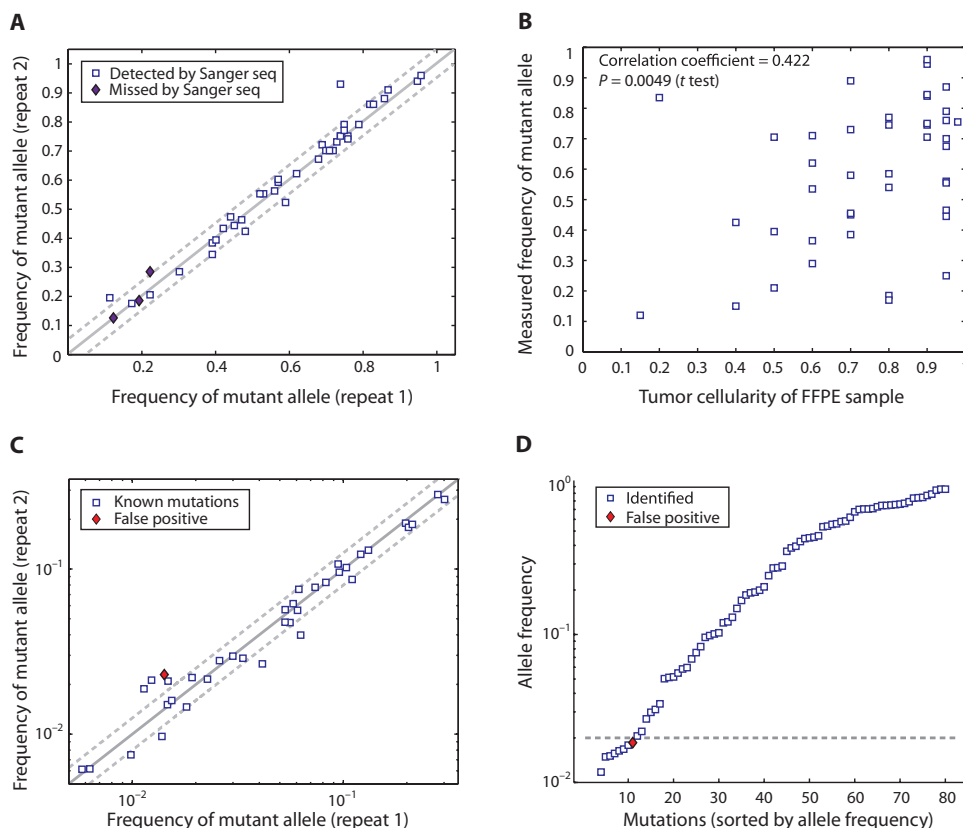


Fig. 2. Identification of mutations in ovarian cancer FFPE samples by TAM-Seq. **(A)** Concordance between duplicate measurements of AFs of mutations identified in fragmented DNA extracted from FFPE samples. The mutation frequency in each library was calculated as the fraction of reads with the mutant (nonreference) base. Solid line indicates equality. Dotted lines indicate a difference in AF of 0.05. **(B)** Correlation of AF with FFPE tumor cellularity. The measured mutant AF (average of both repeats) correlated significantly with the cellularity, estimated from histology (table S3). **(C)** Concordance between duplicate measurements of AFs of mutations identified in a mixture of DNA extracted from different FFPE samples. **(D)** Summary of mutations called in FFPE using TAM-Seq, sorted by increasing AF. Dotted line indicates AF of 2%.

limit of accuracy set by the Poisson distribution for independently segregating molecules. We compared the observed AF to the expected AF for cases where more than six copies of the minor allele were expected. Of 24 such cases, the root mean square (RMS) relative error between the expected and the observed frequency was 14%, with only 2 of 24 cases exhibiting more than 20% discrepancy. For samples with expected minor AF of 0.025, the RMS error was 23% (Fig. 3B).

Noninvasive identification of cancer mutations in plasma circulating DNA

We applied TAM-Seq to directly identify mutations in plasma of cancer patients. We studied a cohort of samples from individuals with HGSOc. These samples were first analyzed for tumor-specific mutations using digital PCR (Supplementary Methods), a method that is highly accurate (2, 3, 7, 37) but requires design and validation of a different assay for every mutation screened and relies on previous identification of mutations in tumor samples from the same patients (2, 3). We initially selected for analysis seven cases that had relatively high levels of circulating mutant *TP53* DNA in the plasma (as assessed by digital PCR). Using the equivalent amount of DNA present in 30

to 120 μ l of plasma, we performed duplicate preamplification reactions for each sample. For all seven patients, *TP53* tumor mutations were identified in the circulating DNA at frequencies of 4% to 44% (Table 1). In one plasma sample collected from an ovarian cancer patient at relapse, we also identified a de novo mutation in the tyrosine kinase domain of *EGFR* (exon 21), at AF of 6% (patient 27, Table 1). We subsequently validated the presence of this mutation in plasma by performing replicate Sanger sequencing reactions of highly diluted template (Supplementary Methods), and 4 of 91 wells that were successfully Sanger-sequenced contained the *EGFR* mutation (fig. S4). We further validated the presence of this mutation by designing a sequence-specific TaqMan probe targeting this mutation and performing digital PCR (Table 1). The mutation was also identified by TAM-Seq in additional plasma collected from the same individual (sample 16, Table 2). This mutation in *EGFR* was not found in the ovarian mass removed by interval debulking surgery 15 months before the blood sample was collected, although the same sample did contain the concomitant *TP53* mutation found in the same patient's plasma, at AF of 85% (patient 27, table S3). We subsequently used TAM-Seq to sequence seven additional samples collected at the time of initial surgery including deposits in right and left ovaries and omentum. The *EGFR* mutation was detected in the two omental samples above the 0.99 confidence margin (fig. S3) at AF of 0.7%, but

was not detected in the six ovarian samples (below the 0.8 confidence margin). Without previous identification in plasma, this mutation would not have been directly identified on screening those samples using high-specificity mutation identification criteria owing to its low AF. In contrast, the *TP53* mutation was identifiable in all biopsy and plasma samples (Fig. 4A). The frequency of mutant alleles in the relapsed tumor could not be directly assessed because a biopsy at relapse was not available.

We validated the TAM-Seq method on a larger panel of plasma samples in which levels of tumor-specific mutations were measured in parallel using patient-specific digital PCR assays. DNA extracted from 62 additional plasma samples collected at different time points from 37 patients with advanced HGSOc was amplified in duplicate (table S6), using DNA present in ~ 0.15 ml of plasma per reaction (range, 0.06 to 0.2 ml). Amplicon libraries were tagged and pooled together for sequencing with libraries prepared from 24 control samples. This generated an average sequencing depth of 650 for 62 plasma samples, sufficient to detect mutations present at AFs of 1% to 2%. Of >1.5 million possible substitutions, 42 mutations were called using the parameters previously optimized for FFPE analysis (table S6).

Fig. 3. Noninvasive identification and quantification of cancer mutations in plasma DNA by TAM-Seq. **(A)** Sampling noise in sequencing of sparse DNA using dilutions of plasma DNA from healthy individuals. CV of triplicate AF readings was calculated for each of the five SNPs in each of the mixes, which had varying numbers of copies of the minor allele (*n*) (blue dots). Bin averages (red diamonds) are the mean CVs calculated for each bin (bin edges denoted by the dotted vertical lines). A linear fit to the log₂ of the mean CV as a function of the log₂ expected copy number was calculated (black line). Two data points, with (*n* = 100, CV = 0.0064) and (*n* = 32, CV = 0.0185), were omitted from the figure for enhanced scaling. Three data points with minor allele copies of <0.8 were omitted from the analysis (*n* = 0.51, CV = 0.62; *n* = 0.41, CV = 0.86; *n* = 0.20, CV = 0.99). **(B)** Expected versus observed frequency of rare alleles in a dilution series of circulating DNA. Mean observed frequency was calculated for each of five SNPs for samples, where expected initial number of minor allele copies was greater than 6. Expected frequencies were calculated on the basis of quantification by digital PCR. Dotted lines represent 20% deviation from the expected frequencies. Inset highlights cases with expected minor AF <0.025. **(C)** Mutations identified in 62 plasma samples from patients with advanced HGSOC using TAM-Seq. AFs are based on digital PCR measurement for confirmed mutations (identified or missed by TAM-Seq), and on TAM-Seq for the false positives called using parameters optimized for analysis

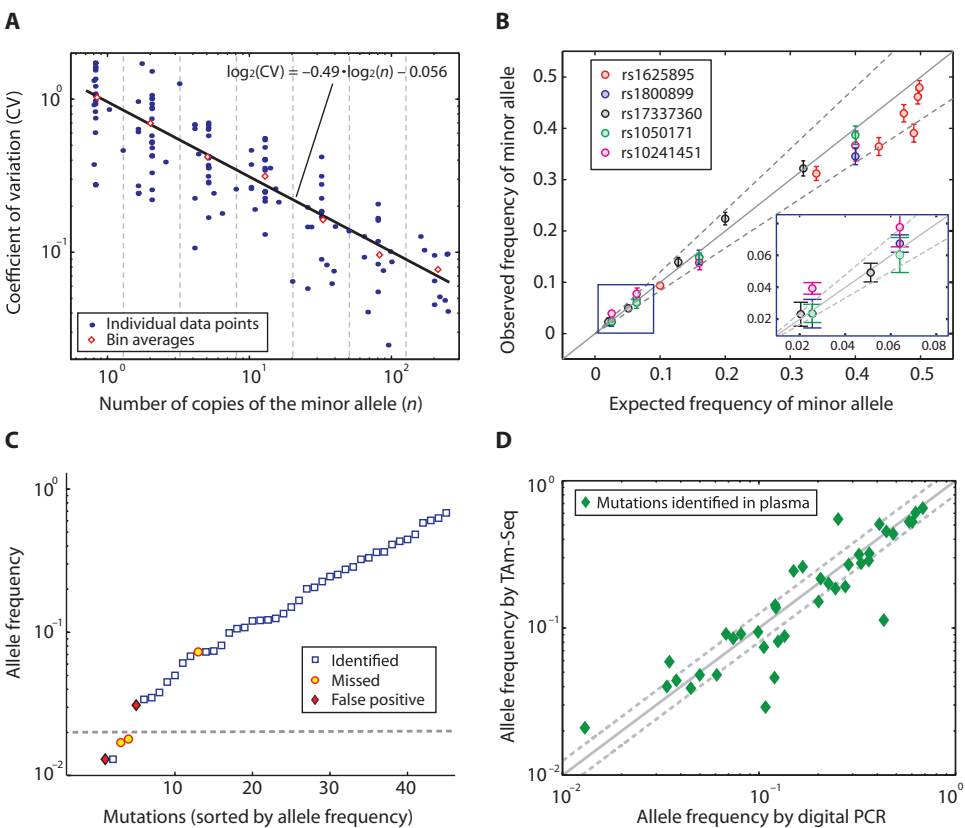


Table 1. Mutations identified by TAM-Seq in plasma samples from seven ovarian cancer patients. TAM-Seq was used to sequence DNA extracted from plasma of subjects with HGSOC (stage III/IV at diagnosis). Plasma was collected when patients presented with relapse disease, before initiation of chemotherapy. For patient 46, DNA from a formalin-fixed, paraffin-

embedded (FFPE) sample was not included in the TAM-Seq set and the mutation was validated in FFPE by Sanger sequencing. CA125 was measured at time of plasma collection. Mean depth of coverage at the mutation locus in the TAM-Seq data was averaged over the repeats (RMS deviation = 850). AF, allele frequency; N, no; Y, yes.

Patient ID	Age at diagnosis	Time elapsed since surgery (months); number of previous lines of chemotherapy	CA125 (U/ml)	Plasma per amplification reaction (μl)	Gene	Mutation and base change (genome build hg19)	Protein change	Detected in FFPE	Mean depth (sequencing reads)	Mean AF using TAM-Seq	Mean AF using digital PCR
8	60	13; 1	2122	50	TP53	17:7577120 C>T	p.R273H	Y	5000	0.09	0.10
12	62	27; 3	365	50	TP53	17:7577579 G>T	p.Y234*	Y	5000	0.10	0.08
14	58	50; 3	260	120	TP53	17:7578212 G>A	p.R213*	Y	5800	0.15	0.12
25	61	9; 1	944	110	TP53	17:7578404 A>T	p.C176S	Y	4800	0.04	0.08
27 [†]	68	15; 1	1051	90	TP53	17:7578262 C>G	p.R196P	Y	7700	0.06	0.14
					EGFR	7:55259437 G>A	p.R832H	N	5700	0.06	0.05
31	64	12; 1	313	30	TP53	17:7578406 C>T	p.R175H	Y	4500	0.44	0.56
46	56	30; 2	1509	30	TP53	17:7578406 C>T	p.R175H	Y	4200	0.23	0.30

*Indicates stop codon. †Both a TP53 and an EGFR mutation were identified in this sample (Fig. 4A).

Table 2. Mutations identified by TAM-Seq in a set of 62 plasma samples from ovarian cancer patients. Forty mutations were identified by TAM-Seq using stringent parameters for mutation calling. Plasma sam-

ples described in this table are distinct from those in Table 1, but patients included overlap. Additional data on patients and mutations are provided in table S6.

Sample number	Plasma volume per amplification reaction (μl)	DNA amount per amplification reaction (ng)	Gene	Protein change	Mean depth (sequencing reads)	Mean AF using TAM-Seq	Mean AF using digital PCR
1	70	0.9	<i>TP53</i>	p.R273C	640	0.260	0.167
2	160	4.2	<i>TP53</i>	p.R248Q	340	0.244	0.150
3	160	5.7	<i>TP53</i>	p.R248Q	640	0.507	0.410
4	120	9.9	<i>TP53</i>	p.R213X	810	0.059	0.035
5	120	1.4	<i>TP53</i>	p.C141Y	680	0.021	0.013
6	120	2.1	<i>TP53</i>	p.C141Y	720	0.044	0.038
7	190	17.9	<i>TP53</i>	p.I195N	800	0.091	0.081
8	160	14.8	<i>TP53</i>	p.R175H	510	0.608	0.627
9	160	10.7	<i>TP53</i>	p.R175H	550	0.526	0.604
10	160	6.1	<i>TP53</i>	p.R175H	530	0.651	0.682
11	160	4.9	<i>TP53</i>	p.R175H	490	0.526	0.581
13	160	2.8	<i>TP53</i>	p.C135R	480	0.039	0.045
14	160	2.5	<i>TP53</i>	p.C135R	610	0.046	0.120
15	160	3.0	<i>TP53</i>	p.C135R	470	0.091	0.068
16 [†]	130	3.7	<i>TP53</i>	p.R196P	1070	0.088	0.135
			<i>EGFR</i>	p.R832H	614	0.048	0.050
17	160	4.2	<i>TP53</i>	p.C176S	580	0.113	0.432
18	160	4.4	<i>TP53</i>	p.C176S	620	0.029	0.108
20	140	5.2	<i>TP53</i>	p.R175H	650	0.201	0.226
21	140	3.6	<i>TP53</i>	p.R175H	650	0.085	0.074
22	140	4.1	<i>TP53</i>	p.R175H	630	0.081	0.125
23	140	3.7	<i>TP53</i>	p.R175H	710	0.074	0.106
24	140	7.1	<i>TP53</i>	p.R175H	760	0.269	0.286
25	130	3.9	<i>TP53</i>	p.R273H	750	0.094	0.099
26	160	5.7	<i>TP53</i>	p.R282W	640	0.048	0.061
27	150	3.6	<i>TP53</i>	p.C141Y	480	0.321	0.364
29	150	9.5	<i>TP53</i>	p.E258K	190	0.548	0.253
31	160	3.6	<i>TP53</i>	p.C135Y	620	0.040	0.034
32	140	2.4	<i>TP53</i>	p.E56X	1480	0.137	0.122
33	160	13.2	<i>TP53</i>	p.K132N	740	0.216	0.206
34	60	5.3	<i>TP53</i>	p.K132N	570	0.151	0.201
36	160	5.8	<i>TP53</i>	p.K132N	620	0.191	0.275
37	160	9.4	<i>TP53</i>	p.K132N	530	0.287	0.362
38	160	10.1	<i>TP53</i>	p.K132N	590	0.275	0.331
39	160	16.4	<i>TP53</i>	p.K132N	700	0.315	0.323
40	160	19.7	<i>TP53</i>	p.K132N	830	0.435	0.482
41	160	15.0	<i>TP53</i>	p.K132N	730	0.452	0.445
42	160	8.5	<i>TP53</i>	p.K132N	560	0.185	0.245
43	150	3.6	<i>TP53</i>	Splicing	680	0.143	0.121
44 [‡]	170	5.2	<i>TP53</i>	p.C238R	1543	0.071	0.073

[†]Both a *TP53* and an *EGFR* mutation were identified in this sample, collected from patient 27 (Table 1), 25 months after initial surgery (Fig. 4A).
amplification in this sample in the initial experiment and was identified successfully in repeat analysis.

[‡]The amplicon containing the mutation failed

Thirty-nine of these matched mutations detected by digital PCR in those samples (Fig. 3C). Three potential false positives were called, at AF of 3.1%, 1.3%, and 0.7% (the latter in a control sample). Using higher-

stringency parameters for mutation identification (Supplementary Methods), we retained only the 39 validated mutations called, with no false positives (Table 2).

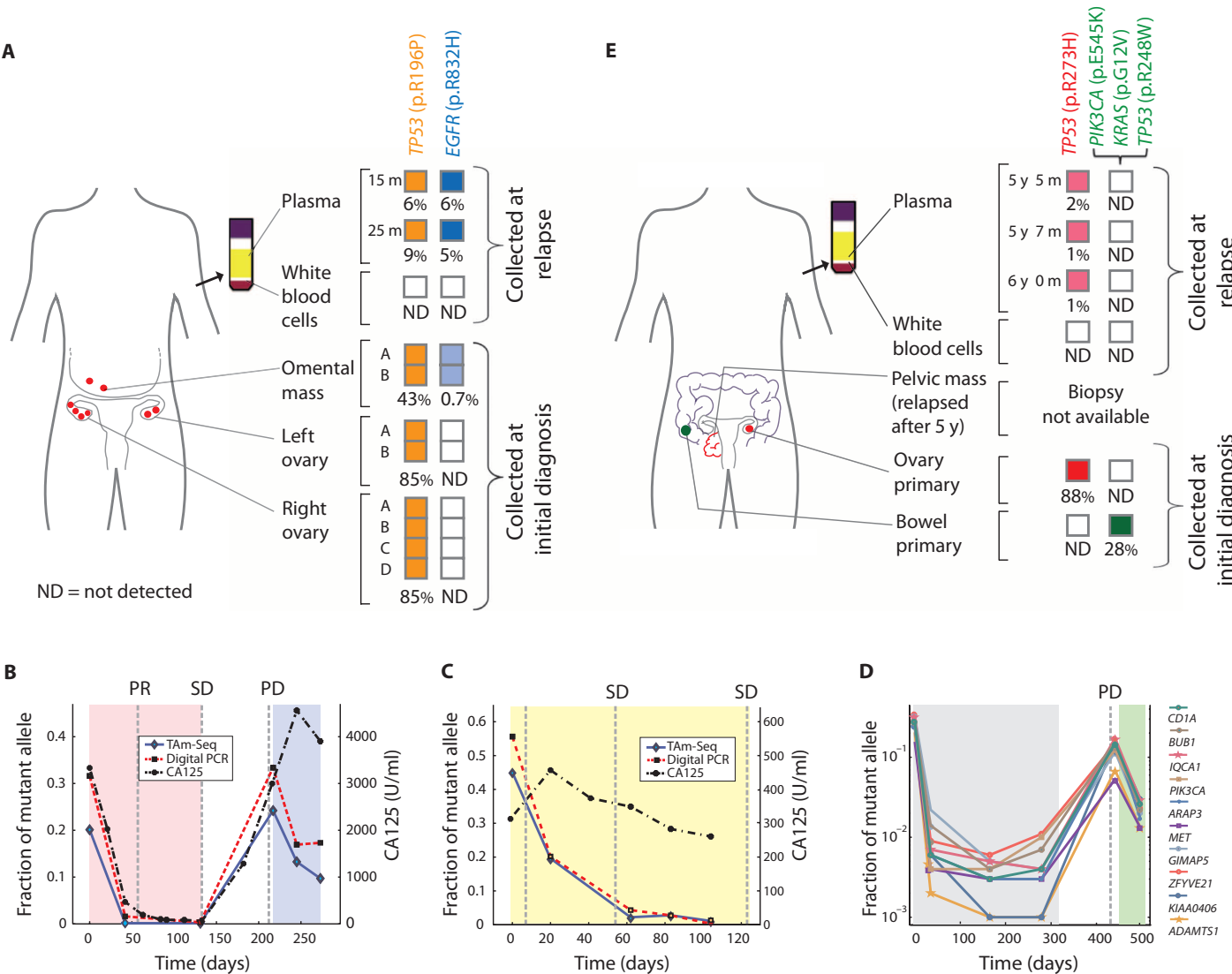


Fig. 4. Clinically relevant applications of plasma DNA sequencing using TAM-Seq. **(A)** Retrospective analysis by TAM-Seq of plasma samples collected during patient follow-up and biopsy specimens collected at initial surgery. We identified a mutation in exon 21 of *EGFR* (dark blue boxes) in two separate plasma samples, collected 15 and 25 months after initial surgery from patient 27 (Tables 1 and 2). This mutation was not directly identified in eight tumor biopsy specimens collected at the time of initial surgery (two from omental mass, two from left ovary, and four from right ovary). Having identified the mutation in the plasma samples, we examined this mutation using the lower-specificity criteria defined for mutation detection (Supplementary Methods) and detected the mutation in the two specimens that had been collected from the omentum at the time of surgery (light blue boxes) but not in the six ovarian specimens. A mutation in *TP53* was identified in all tumor and plasma samples collected from this patient (Tables 1 and 2 and table S3), but not in white blood cells (buffy coat). Percentages indicate mutant AFs. Empty boxes and “ND” indicate samples where a mutation was not identified or detected (below 0.8 confidence margin). **(B)** Monitoring frequency

of mutant DNA in plasma of an ovarian cancer patient (patient 46) over time using TAM-Seq and digital PCR. TAM-Seq results are reported as the mean frequency of duplicate analyses. Parallel data are shown for digital PCR and serum CA125. Shaded regions indicate periods of chemotherapy, and vertical dashed lines indicate radiological assessment of patient responses: PR, partial response; SD, stable disease; PD, progressive disease. **(C)** Monitoring frequency of mutant DNA in plasma of an ovarian cancer patient (patient 31) over time. **(D)** Dynamics of 10 tumor-specific mutations in plasma of a breast cancer patient (not included in the other sets of samples analyzed). **(E)** Retrospective analysis of samples from synchronous primary tumors (bowel and ovarian) collected at the time of initial surgery and three plasma samples collected at relapse. In primary tumors from this patient (not included in the other sets of samples analyzed), a *TP53* mutation was identified in the ovarian cancer (red box), and mutations in *PIK3CA*, *KRAS*, and *TP53* were identified in the bowel cancer (green box). At relapse, a biopsy was not performed on the pelvic mass. The *TP53* mutation that was identified in the ovarian primary tumor (p.R273H) was detected in plasma, whereas the bowel-associated mutations were not detected.

Of 40 point mutations detected at AF >2% by digital PCR, 38 (95%) were identified by TAm-Seq in a single experiment (Fig. 3C). One additional mutation was located in an amplicon that failed in that sample and was identified in repeated analysis; the other was likely missed by TAm-Seq owing to sampling noise, because it was found in one of the duplicate preamplified libraries but not the other (table S6). One of three mutations detected by digital PCR at 1% < AF < 2% was identified by TAm-Seq (Fig. 3C). Eleven additional point mutations detected by digital PCR at AF <1% were not detected by TAm-Seq at these settings. TAm-Seq and digital PCR measurements of AF had excellent agreement, with correlation coefficient of 0.90, increasing to 0.97 when discarding the two strongest outliers (Fig. 3D). Thus, we screened 62 samples across sizeable genomic stretches, using minute amounts of plasma DNA (median, 4 ng), and obtained 97.5% sensitivity with PPV of 100% for identifying mutations at AF >2% in plasma by TAm-Seq. Using parameters optimized for FFPE samples, one potential false positive was called at AF >2%, reducing the PPV to 97.5% (Table 3).

Monitoring levels of ctDNA

Various methods have been suggested to monitor changes in mutation load in plasma. These can have enhanced sensitivity compared to TAm-Seq for tracking individual mutations, but require design of personalized assays (3, 18, 19). None of these methods have been widely adopted. We therefore applied TAm-Seq as a generic tool to measure changes in the frequency of ctDNA over time. We studied serial plasma

samples collected during follow-up and treatment of two patients with relapsed HGSOc, collected during 104 and 273 days of follow-up and treatment, respectively. Frequencies of mutant *TP53* alleles were measured by TAm-Seq and in parallel by digital PCR using a mutation-specific probe. The two methods of quantification had excellent agreement. Mutant AFs in plasma of ovarian cancer patients reflected well the clinical course of the disease compared to the serum marker CA125, showed marked decrease when systemic treatment was initiated, and increased in parallel to disease progression. In the first case (Fig. 4B), a 56-year-old woman with relapsed ovarian cancer (patient 46) was treated with fourth-line carboplatin + paclitaxel chemotherapy for six cycles (pink-shaded region). Radiology showed partial response on mid-treatment computed tomography (CT) scan. End-of-treatment CT showed stable disease. Twelve weeks from the end of her fourth-line treatment, the patient developed progressive disease. The patient then initiated fifth-line chemotherapy with liposomal doxorubicin (purple-shaded region). In the second case (Fig. 4C), a 64-year-old woman with relapsed ovarian cancer (patient 31) was treated with second-line ECX (epirubicin, cisplatin, and capecitabine) chemotherapy for six cycles. Radiology showed stable disease on mid- and end-of-treatment CT scans. The patient then remained off treatment, until she progressed 3 months later.

TAm-Seq can be flexibly adapted to sequence different genomic regions by designing primers to amplify regions of interest. We used this capability to study dynamics of multiple mutations in parallel.

Table 3. Summary of mutations identified in 69 plasma samples of ovarian cancer patients. Samples were analyzed by TAm-Seq and in parallel by digital PCR. Using parameters optimized for plasma DNA, false-positive calls were lost, whereas all confirmed calls were retained, resulting in specificity and PPV of 100%.

First set of plasma samples	
Plasma samples analyzed	7
Point mutations originally detected by digital PCR, using patient-specific assays targeting mutations identified in tumor samples	7
Point mutations identified directly in plasma by TAm-Seq	8
De novo mutations identified by TAm-Seq only, subsequently confirmed by digital PCR	1
Second set of plasma samples	
Plasma samples analyzed	62
Point mutations detected by digital PCR at AF >2%	40
Point mutations with AF >2% (by digital PCR) identified by TAm-Seq	39
Point mutations missed by TAm-Seq due to sampling error	1
Sensitivity of TAm-Seq for identifying mutations at AF >2%	97.5%
PPV of mutations called by TAm-Seq with AF >2%	97.5%*
ctDNA in ovarian cancer	
Advanced ovarian cancer patients in both sets†	38
Patients where TAm-Seq identified cancer mutations	20

*One unconfirmed substitution was called at AF >2% using parameters optimized for FFPE material. †The first set included 7 patients (Table 1), and the second set included 37 patients (table S6), 6 of whom overlap.

Whole-genome sequencing of tumor material was used to identify tumor mutations in a patient with metastatic breast cancer undergoing two phases of chemotherapy. Ten mutations were selected, and short amplicons (<120 bp) were designed to cover the mutation loci (table S7). Serial plasma samples were collected over the course of 497 days, both before and after treatment. We performed TAm-Seq in duplicate, using DNA from 0.08 ml of plasma per amplification, and tracked dynamics of all mutations in parallel (Fig. 4D). The patient was treated with single-agent epirubicin (gray-shaded region). After 4 months off treatment, a CT scan showed progressive disease and the patient commenced further treatment with paclitaxel chemotherapy. The 10 mutations followed a common pattern of sharp decline in AF upon onset of therapy and an increase in AF upon disease progression after termination of therapy (Fig. 4D).

Finally, we used TAm-Seq to study plasma from a patient who had a history of two synchronous primary cancers, bowel and ovarian, which were resected simultaneously. After a 5-year remission, a pelvic mass of uncertain origin was detected. A biopsy was considered to guide selection of therapy but was not performed owing to risk of complications and comorbidities. The patient commenced empirically on an ovarian cancer chemotherapy regimen, to which she responded. Retrospective analysis by TAm-Seq of FFPE from the primary tumors collected at initial surgery, and three plasma samples collected serially at the time of relapse (5 years and 5 months, 5 years and 7 months, and 6 years after initial surgery), showed that the patient's plasma at relapse contained the *TP53* (p.R273H) mutation identified in the ovarian primary tumor (exceeding the 0.98, 0.93, and 0.97 confidence margins, respectively), but not the *PIK3CA* (p.E545K), *KRAS* (p.G12V), or *TP53* (p.R248W) mutations identified in the primary bowel cancer (below the 0.8 confidence margin) (Fig. 4E). Had these results been available, uncertainty and treatment delays may have been avoided, as well as the risk of prescribing chemotherapy for an inappropriate tumor site. An alternative possible outcome may have involved a finding of the *PIK3CA* or *KRAS* mutations (present in the primary bowel cancer) in the patient's plasma at the time of relapse. Such a finding, if available to clinicians at the time, may not only have led to alternate chemotherapy being offered but may have also opened the possibility of enrolment into a trial for targeted therapy with mammalian target of rapamycin (mTOR), phosphatidylinositol 3-kinase (PI3K), or mitogen-activated protein kinase kinase (MEK) inhibitors (11).

DISCUSSION

Detection of rare mutations in circulating DNA has long been pursued owing to its potentially transformative impact on cancer diagnosis and management. Important progress has been made using sequence-specific assays that target predefined mutations and that detect extremely rare alleles. Assays such as PCR (6, 7), ligation (5), and primer extension/mass spectrometry (27) can identify specific, predefined mutations in plasma samples. Enhanced detection down to 1 mutant allele in 10,000 or more wild-type alleles can be obtained using a variety of methods, such as peptide nucleic acid and primer extension ("PPEM") (38), ligation followed by quantitative PCR ("LigAmp") (39), bead-based digital PCR in emulsions ("BEAMing") (2, 3), microfluidic-based (7) or droplet-based digital PCR (40), or microinsertion/deletion/indel-activated pyrophosphorolysis ("MAP") (29). Nonetheless, identification of rare mutations in tumor suppressor genes such as *TP53*,

which are widely mutated in cancers but lack a well-defined hotspot region, remains an elusive goal.

In patients with advanced cancers, mutant alleles can reach a sizeable fraction of DNA. For example, Dukes' D colorectal cancers have median 8% mutant AF (2). Screening of entire genes for mutations would therefore be useful for some applications, even if analytical selectivity is limited to a few percent. Advances in massively parallel sequencing make new approaches possible. These have largely focused on large-scale analyses, including whole-genome or whole-exome sequencing (41). This generates a large amount of data on genomic regions that do not, at present, inform clinical decisions. Moreover, the depth of coverage for clinically significant loci is not sufficient to detect changes that occur at low frequency (<5%). Such approaches have recently been complemented by methods for examination of individual amplicons at great depth (30).

The intermediate scale of sequencing is most likely to have immediate impact on clinical genomics. Targeted sequencing has been applied for tumor DNA (34, 35) and cyst fluid (33) to detect mutations down to 5% AF, but has not been applied for analysis of circulating tumor nucleic acids. Here, we demonstrate noninvasive identification of mutant alleles in plasma, at AFs as low as 2%, by targeted deep sequencing of circulating DNA. Our TAm-Seq method uses a combination of short amplicons, two-step amplification, sample barcodes, and high-throughput PCR. Because the amplicons are short, this method effectively amplifies even small amounts of fragmented DNA such as are present in circulating DNA. The two-step amplification permits extensive primer multiplexing that enables the amplification and sequencing of sizeable genomic regions by tiling short amplicons without loss of fidelity or efficiency. Duplicate sequencing of each sample is used to avoid false positives stemming from PCR errors. Sample barcodes and high-throughput PCR reduce the per-sample costs to a range where this may be widely applicable. Preparing TAm-Seq libraries for sequencing from 48 samples takes less than 24 hours and involves only a few hours of hands-on time. New platforms for massively parallel sequencing allow for fast turnaround times, which make this approach practical in a clinical setting.

The sensitivity presently achieved can provide useful diagnostic information in certain advanced cancers. We studied a cohort of subjects with advanced HGSOc in which the tumor suppressor gene *TP53* is a driver mutation (20). Of the 69 plasma samples collected from 38 different individuals with advanced HGSOc, we identified mutations in *TP53* in 46 samples (67%) from 20 of the cases (53%). In contrast, a previous study using a ligase detection reaction with bespoke primers found mutated *TP53* sequences in plasma for only 30% of advanced ovarian cancer patients (5), and a study using single-strand conformation polymorphism found no ctDNA in preoperative plasma samples from high-grade serous cancer patients (42).

Targeted agents, such as inhibitors of poly(adenosine diphosphate-ribose) polymerase (PARP), or tyrosine kinase inhibitors targeting epidermal growth factor receptor (EGFR), may be applicable for systemic treatment of advanced HGSOc (8, 10, 22). In a recent study of 203 HGSOc tumors, *EGFR* was found to be the most frequently mutated oncogene and was mutated in nearly 10% of cases (10). In one case, we identified in plasma a de novo mutation in the tyrosine kinase domain (exon 21) of *EGFR*, located 26 amino acids upstream of the L858R activating mutation widely documented for lung cancer. In a subset of tumor samples collected from the same patient 15 months earlier, this mutation was detected at AF of 0.7%, but could not have

been identified by analysis of those samples alone without previous knowledge of the mutation identified in plasma (Fig. 4A). In a clinical setting, identification of such a mutation could potentially guide treatment with alternative molecularly targeted therapy (10). Current clinical recommendations in lung adenocarcinoma suggest mutation assessment in exons 18 to 21 of *EGFR* (a region of ~560 bp) in the tumor tissue to identify patients eligible for treatment with gefitinib or erlotinib (9). Using a commercial PCR-based in vitro diagnostic kit (Qiagen), 28 different *EGFR* variants can be assayed (not including the mutation we identified), but the sample needs to be subdivided into seven different reactions. When sample is limited or mutant alleles are rare, this could introduce sampling errors.

Using standard amplification primers tailored to the mutation loci, we also used TAm-Seq to monitor the dynamics of 10 mutations in plasma DNA of a single patient with metastatic breast cancer, using minute amounts of input DNA. Previous studies have followed up to two mutations in any individual patient (3, 19). Tracking multiple mutations can provide insight into clonal evolution and, at the same time, increases the robustness for tumor monitoring by compensating for effects of sampling noise or mutational drift. For example, if a patient has only five copies of a mutant allele per milliliter of plasma (on average), there is a 37% probability that this mutation will not be present in a 0.2-ml sample, and even a perfect assay will fail to detect residual tumor, whereas a method that measures multiple mutations in parallel can have a low likelihood of a false-negative result even if the detection rate for each mutation is less than 50%.

A current limitation of TAm-Seq is the detection limit compared to assays that target individual loci (2, 3, 7, 40), which have been shown to detect two to three orders of magnitude lower frequencies. Our approach may be sufficient for analyzing plasma from patients with certain advanced cancers, but further improvement may be necessary before this method can be more widely used in the clinic. Higher read depth or fidelity, additional replicates, or improved algorithms could allow for enhanced mutation detection without change to protocols. An alternative strategy is through rare allele enrichment, for example, by combining TAm-Seq with protocols such as COLD-PCR (co-amplification at lower denaturation temperature PCR) (31).

Previously proposed methods for personalized monitoring of tumor dynamics relied on expensive custom-designed probes (3) or identification of rearrangements using whole-genome sequencing (18, 19). These have better analytical sensitivity than currently achieved by TAm-Seq, but are difficult to implement on a routine basis. TAm-Seq strikes a balance between sensitivity and ease of use and could facilitate study and application of circulating DNA. Using TAm-Seq, we identified cancer mutations in the plasma of most advanced ovarian cancer patients and tracked dynamics of *TP53* mutations without requiring any specially designed probes. In summary, TAm-Seq is a flexible and cost-effective platform for applications in noninvasive cancer genomics and diagnostics. We have shown that this method can be used for high-throughput sequencing of plasma samples to identify and monitor levels of multiple cancer mutations in circulating DNA. This could also be applied to screen for rare mutations in a variety of heterogeneous sample types such as low-cellularity tumor specimens, cytological samples, or circulating tumor cells (16). With further developments, this and derivative methods may be applied in molecular screening for earlier detection or for differential diagnosis of cancer from benign masses. For genetic analysis of FFPE or small biopsy samples, TAm-Seq can be applied as is, as a cost-effective clinical aid.

MATERIALS AND METHODS

Sample collection

FFPE blocks were obtained from the pathology archives at Addenbrooke's Hospital (Cambridge, UK). Plasma samples were collected upon disease relapse, before and during chemotherapy treatment. Sample collection for this study was approved by Cambridgeshire Research Ethics Committee (REC 08/H0306/61 and 07/Q0106/63). Peripheral blood samples were collected into EDTA tubes and centrifuged at 820g for 10 min within 1 hour of collection to limit degradation of cell-free DNA and leukocyte lysis. Aliquots (1 ml) of plasma were centrifuged in a bench-top microfuge at 14,000 rpm for 10 min. The supernatant was transferred to sterile 1.5-ml tubes and stored at -80°C before extraction.

Extraction of DNA from FFPE and blood plasma

Paraffin blocks were cut as 8- μ m sections on plain glass slides. Targeted regions for sampling were marked on adjacent hematoxylin and eosin sections by the study pathologist and recovered by scrape macrodissection. Between 3 and 20 sections were macrodissected depending on the tissue sample's size. DNA from FFPE sections was extracted with QIAamp DNA FFPE Tissue Kit (Qiagen) according to the manufacturer's instructions.

Circulating DNA was extracted from between 0.85 and 2.2 ml of plasma with the QIAamp Circulating Nucleic Acid kit (Qiagen), following the manufacturer's instructions, and with the QIAvac 24 Plus vacuum manifold. Carrier RNA was added to ACL lysis buffer to enhance binding of nucleic acids to the QIAamp membrane with the aim to enhance yields.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/4/136/136ra68/DC1

Methods

Fig. S1. PCR strategy and primer design.

Fig. S2. Sanger traces for mutations identified by tagged-amplicon sequencing.

Fig. S3. Background frequencies and detection limits for base substitutions.

Fig. S4. Replicate dilute Sanger sequencing of a mutation identified in plasma.

Table S1. Target-specific primers.

Table S2. Unique sequencing barcodes.

Table S3. Mutations identified in FFPE samples.

Table S4. SNPs identified in circulating DNA from two plasma control samples.

Table S5. Frequency of SNP alleles in dilution series of DNA from control plasma.

Table S6. Additional data for Table 2 for mutations identified in plasma samples.

Table S7. Mutations and amplicons studied in one breast cancer patient.

REFERENCES AND NOTES

1. E. Gormally, E. Caboux, P. Vineis, P. Hainaut, Circulating free DNA in plasma or serum as biomarker of carcinogenesis: Practical aspects and biological significance. *Mutat. Res.* **635**, 105–117 (2007).
2. F. Diehl, M. Li, D. Dressman, Y. He, D. Shen, S. Szabo, L. A. Diaz Jr., S. N. Goodman, K. A. David, H. Juhl, K. W. Kinzler, B. Vogelstein, Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16368–16373 (2005).
3. F. Diehl, K. Schmidt, M. A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokoll, S. A. Szabo, K. W. Kinzler, B. Vogelstein, L. A. Diaz Jr., Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).
4. H. Schwarzenbach, D. S. Hoon, K. Pantel, Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).
5. E. M. Swisher, M. Wollan, S. M. Mahtani, J. B. Willner, R. Garcia, B. A. Goff, M. C. King, Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *Am. J. Obstet. Gynecol.* **193**, 662–667 (2005).

6. R. E. Board, A. M. Wardley, J. M. Dixon, A. C. Armstrong, S. Howell, L. Renshaw, E. Donald, A. Greystoke, M. Ranson, A. Hughes, C. Dive, Detection of *PIK3CA* mutations in circulating free DNA in patients with breast cancer. *Breast Cancer Res. Treat.* **120**, 461–467 (2010).
7. T. K. Yung, K. C. Chan, T. S. Mok, J. Tong, K. F. To, Y. M. Lo, Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. *Clin. Cancer Res.* **15**, 2076–2084 (2009).
8. S. Banerjee, S. Kaye, The role of targeted therapy in ovarian cancer. *Eur. J. Cancer* **47** (Suppl. 3), S116–S130 (2011).
9. V. A. Keedy, S. Temin, M. R. Somerfield, M. B. Beasley, D. H. Johnson, L. M. McShane, D. T. Milton, J. R. Strawn, H. A. Wakelee, G. Giaccone, American Society of Clinical Oncology provisional clinical opinion: Epidermal growth factor receptor (*EGFR*) mutation testing for patients with advanced non-small-cell lung cancer considering first-line *EGFR* tyrosine kinase inhibitor therapy. *J. Clin. Oncol.* **29**, 2121–2127 (2011).
10. U. A. Matulonis, M. Hirsch, E. Palescandolo, E. Kim, J. Liu, P. van Hummelen, L. MacConaill, R. Drapkin, W. C. Hahn, High throughput interrogation of somatic mutations in high grade serous cancer of the ovary. *PLoS One* **6**, e24433 (2011).
11. J. A. Engelman, L. Chen, X. Tan, K. Crosby, A. R. Guimaraes, R. Upadhyay, M. Maira, K. McNamara, S. A. Perera, Y. Song, L. R. Chirieac, R. Kaur, A. Lightbown, J. Simendinger, T. Li, R. F. Padera, C. Garcia-Echeverria, R. Weissleder, U. Mahmood, L. C. Cantley, K. K. Wong, Effective use of PI3K and MEK inhibitors to treat mutant *Kras* G12D and *PIK3CA* H1047R murine lung cancers. *Nat. Med.* **14**, 1351–1356 (2008).
12. P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. A. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, C. A. Griffin, J. Burton, H. Swerdlow, M. A. Taylor, M. R. Stratton, C. Iacobuzio-Donahue, P. A. Futreal, The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
13. S. P. Shah, R. D. Morin, J. Khattri, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, S. Aparicio, Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
14. M. Inukai, S. Toyooka, S. Ito, H. Asano, S. Ichihara, J. Soh, H. Suehisa, M. Ouchida, K. Aoe, M. Aoe, K. Kiura, N. Shimizu, H. Date, Presence of *epidermal growth factor receptor gene* T790M mutation as a minor clone in non-small cell lung cancer. *Cancer Res.* **66**, 7854–7858 (2006).
15. S. L. Edwards, R. Brough, C. J. Lord, R. Natrajan, R. Vatcheva, D. A. Levine, J. Boyd, J. S. Reis-Filho, A. Ashworth, Resistance to therapy caused by intragenic deletion in *BRCA2*. *Nature* **451**, 1111–1115 (2008).
16. S. Maheswaran, L. V. Sequist, S. Nagrath, L. Ulkus, B. Brannigan, C. V. Collura, E. Inersa, S. Diederichs, A. J. Iafrate, D. W. Bell, S. Digumarthy, A. Muzikansky, D. Irimia, J. Settleman, R. G. Tompkins, T. J. Lynch, M. Toner, D. A. Haber, Detection of mutations in *EGFR* in circulating lung-cancer cells. *N. Engl. J. Med.* **359**, 366–377 (2008).
17. B. Norquist, K. A. Wurz, C. C. Pennil, R. Garcia, J. Gross, W. Sakai, B. Y. Karlan, T. Taniguchi, E. M. Swisher, Secondary somatic mutations restoring *BRCA1/2* predict chemotherapy resistance in hereditary ovarian carcinomas. *J. Clin. Oncol.* **29**, 3008–3015 (2011).
18. R. J. Leary, I. Kinde, F. Diehl, K. Schmidt, C. Clouser, C. Duncan, A. Antipova, C. Lee, K. McKernan, F. M. De La Vega, K. W. Kinzler, B. Vogelstein, L. A. Diaz Jr., V. E. Velculescu, Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
19. D. J. McBride, A. K. Orpana, C. Sotiriou, H. Joensuu, P. J. Stephens, L. J. Mudie, E. Hämläinen, L. A. Stebbings, L. C. Andersson, A. M. Flanagan, V. Durbecq, M. Ignatiadis, O. Kallioniemi, C. A. Heckman, K. Alitalo, H. Edgren, P. A. Futreal, M. R. Stratton, P. J. Campbell, Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer* **49**, 1062–1069 (2010).
20. A. A. Ahmed, D. Etemadmoghadam, J. Temple, A. G. Lynch, M. Riad, R. Sharma, C. Stewart, S. Fereday, C. Caldas, A. Defazio, D. Bowtell, J. D. Brenton, Driver mutations in *TP53* are ubiquitous in high grade serous carcinoma of the ovary. *J. Pathol.* **221**, 49–56 (2010).
21. Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
22. R. C. Bast Jr., Molecular approaches to personalizing management of ovarian cancer. *Ann. Oncol.* **22** (Suppl. 8), viii5–viii15 (2011).
23. K. C. Chan, S. F. Leung, S. W. Yeung, A. T. Chan, Y. M. Lo, Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **14**, 4141–4145 (2008).
24. H. C. Fan, Y. J. Blumenfeld, U. Chitkara, L. Hudgins, S. R. Quake, Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin. Chem.* **56**, 1279–1286 (2010).
25. Y. M. Lo, K. C. Chan, H. Sun, E. Z. Chen, P. Jiang, F. M. Lun, Y. W. Zheng, T. Y. Leung, T. K. Lau, C. R. Cantor, R. W. Chiu, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
26. R. W. Chiu, R. Akolekar, Y. W. Zheng, T. Y. Leung, H. Sun, K. C. Chan, F. M. Lun, A. T. Go, E. T. Lau, W. W. To, W. C. Leung, R. Y. Tang, S. K. Au-Yeung, H. Lam, Y. Y. Kung, X. Zhang, J. M. van Vugt, R. Minekawa, M. H. Tang, J. Wang, C. B. Oudejans, T. K. Lau, K. H. Nicolaides, Y. M. Lo, Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: Large scale validity study. *BMJ* **342**, c7401 (2011).
27. Y. M. Lo, N. B. Tsui, R. W. Chiu, T. K. Lau, T. N. Leung, M. M. Heung, A. Gerovassili, Y. Jin, K. H. Nicolaides, C. R. Cantor, C. Ding, Plasma placental RNA allelic ratio permits non-invasive prenatal chromosomal aneuploidy detection. *Nat. Med.* **13**, 218–223 (2007).
28. Y. M. Lo, F. M. Lun, K. C. Chan, N. B. Tsui, K. C. Chong, T. K. Lau, T. Y. Leung, B. C. Zee, C. R. Cantor, R. W. Chiu, Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 13116–13121 (2007).
29. Z. Chen, J. Feng, C. H. Buzin, Q. Liu, L. Weiss, K. Kernstine, G. Somlo, S. S. Sommer, Analysis of cancer mutation signatures in blood by a novel ultra-sensitive assay: Monitoring of therapy or recurrence in non-metastatic breast cancer. *PLoS One* **4**, e7220 (2009).
30. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9530–9535 (2011).
31. J. Li, L. Wang, H. Mamon, M. H. Kulke, R. Berbeco, G. M. Makrigiorgos, Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat. Med.* **14**, 579–584 (2008).
32. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, P. A. Futreal, COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
33. J. Wu, H. Matthaei, A. Maitra, M. Dal Molin, L. D. Wood, J. R. Eshleman, M. Goggins, M. I. Canto, R. D. Schulick, B. H. Edil, C. L. Wolfgang, A. P. Klein, L. A. Diaz Jr., P. J. Allen, C. M. Schmidt, K. W. Kinzler, N. Papadopoulos, R. H. Hruban, B. Vogelstein, Recurrent *GNAS* mutations define an unexpected pathway for pancreatic cyst development. *Sci. Transl. Med.* **3**, 92ra66 (2011).
34. O. Harismendy, R. B. Schwab, L. Bao, J. Olson, S. Rozenzhak, S. K. Kotsopoulos, S. Pond, B. Crain, M. S. Chee, K. Messer, D. R. Link, K. A. Frazer, Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol.* **12**, R124 (2011).
35. N. Wagle, M. F. Berger, M. J. Davis, B. Blumenstiel, M. DeFelice, P. Pochanard, M. Ducar, P. Van Hummelen, L. E. MacConaill, W. C. Hahn, M. Meyerson, S. B. Gabriel, L. A. Garraway, High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82 (2012).
36. T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
37. B. Vogelstein, K. W. Kinzler, Digital PCR. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9236–9241 (1999).
38. X. Sun, K. Hung, L. Wu, D. Sidransky, B. Guo, Detection of tumor mutations in the presence of excess amounts of normal DNA. *Nat. Biotechnol.* **20**, 186–189 (2002).
39. C. Shi, S. H. Eshleman, D. Jones, N. Fukushima, L. Hua, A. R. Parker, C. J. Yeo, R. H. Hruban, M. G. Goggins, J. R. Eshleman, LigAmp for sensitive detection of single-nucleotide differences. *Nat. Methods* **1**, 141–147 (2004).
40. L. B. Pinheiro, V. A. Coleman, C. M. Hindson, J. Herrmann, B. J. Hindson, S. Bhat, K. R. Emslie, Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Anal. Chem.* **84**, 1003–1011 (2012).
41. M. Meyerson, S. Gabriel, G. Getz, Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
42. J. Otsuka, T. Okuda, A. Sekizawa, S. Amemiya, H. Saito, T. Okai, M. Kushima, Detection of p53 mutations in the plasma DNA of patients with ovarian cancer. *Int. J. Gynecol. Cancer* **14**, 459–464 (2004).

Acknowledgments: We thank H. Biggs, C. Hodgkin, S. Richardson, and L. Jones for assistance in sample collection, S. Aldridge for assistance in genomic analysis, and S. Tavaré, B. Davis, and M. Dunning for assistance in data analysis. **Funding:** We acknowledge the support of Cancer Research UK, the University of Cambridge, National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge Experimental Cancer Medicine Centre, and Hutchison Whampoa Limited. C.P. was supported in part by the Academy of Medical Sciences, Wellcome Trust, British Heart Foundation, and Arthritis Research UK. **Author contributions:** T.F., M.M., C.P., D.G., D.W.Y.T., C.C., J.D.B., and N.R. designed the study. T.F., M.M., D.W.Y.T., F.K., J.H., A.P.M.,

and N.R. developed methods. T.F., D.G., D.W.Y.T., A.M.P., and S.-J.D. collected the data. T.F., M.M., and N.R. analyzed TAm-Seq data. C.P., S.-J.D., C.C., and J.D.B. designed clinical studies and collected samples and clinical data. M.J.-L. performed pathological analysis. D.B. contributed sequencing data. T.F., M.M., C.P., D.G., D.W.Y.T., J.H., A.P.M., J.D.B., and N.R. interpreted the data. T.F., M.M., and N.R. wrote the manuscript with assistance from C.P., D.G., D.W.Y.T., A.P.M., J.D.B., and other authors. All authors approved the final manuscript. **Competing interests:** A.P.M. and F.K. hold equity in Fluidigm and may stand to gain by publication of these findings. D.B. and F.K. hold equity in Illumina and may stand to gain by publication of these findings.

Submitted 24 October 2011

Accepted 18 April 2012

Published 30 May 2012

10.1126/scitranslmed.3003726

Citation: T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Y. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld, Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012).

Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA

Muhammed Murtaza^{1*}, Sarah-Jane Dawson^{1,2*}, Dana W. Y. Tsui^{1*}, Davina Gale¹, Tim Forshe¹, Anna M. Piskorz¹, Christine Parkinson^{1,2}, Suet-Feung Chin¹, Zoya Kingsbury³, Alvin S. C. Wong⁴, Francesco Marass¹, Sean Humphray³, James Hadfield¹, David Bentley³, Tan Min Chin^{4,5}, James D. Brenton^{1,2,6}, Carlos Caldas^{1,2,6} & Nitzan Rosenfeld¹

Cancers acquire resistance to systemic treatment as a result of clonal evolution and selection^{1,2}. Repeat biopsies to study genomic evolution as a result of therapy are difficult, invasive and may be confounded by intra-tumour heterogeneity^{3,4}. Recent studies have shown that genomic alterations in solid cancers can be characterized by massively parallel sequencing of circulating cell-free tumour DNA released from cancer cells into plasma, representing a non-invasive liquid biopsy^{5–7}. Here we report sequencing of cancer exomes in serial plasma samples to track genomic evolution of metastatic cancers in response to therapy. Six patients with advanced breast, ovarian and lung cancers were followed over 1–2 years. For each case, exome sequencing was performed on 2–5 plasma samples (19 in total) spanning multiple courses of treatment, at selected time points when the allele fraction of tumour mutations in plasma was high, allowing improved sensitivity. For two cases, synchronous biopsies were also analysed, confirming genome-wide representation of the tumour genome in plasma. Quantification of allele fractions in plasma identified increased representation of mutant alleles in association with emergence of therapy resistance. These included an activating mutation in *PIK3CA* (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha) following treatment with paclitaxel⁸; a truncating mutation in *RBI* (retinoblastoma 1) following treatment with cisplatin⁹; a truncating mutation in *MED1* (mediator complex subunit 1) following treatment with tamoxifen and trastuzumab^{10,11}, and following subsequent treatment with lapatinib^{12,13}, a splicing mutation in *GAS6* (growth arrest-specific 6) in the same patient; and a resistance-conferring mutation in *EGFR* (epidermal growth factor receptor; T790M) following treatment with gefitinib¹⁴. These results establish proof of principle that exome-wide analysis of circulating tumour DNA could complement current invasive biopsy approaches to identify mutations associated with acquired drug resistance in advanced cancers. Serial analysis of cancer genomes in plasma constitutes a new paradigm for the study of clonal evolution in human cancers.

Serial sampling of the tumour genome is required to identify the mutational mechanisms underlying drug resistance². Serial tumour biopsies are invasive and often unattainable. Tumours are heterogeneous and continuously evolve, and even if several biopsies are obtained, these are limited both spatially and temporally. Analysis of isolated circulating tumour cells (CTCs) has been proposed, but circulating tumour DNA (ctDNA) is more accessible and easier to process¹⁵. Previous studies of tumour mutations in plasma have analysed individual loci, genes or structural variants to quantify tumour burden and to detect previously-characterized resistance-conferring mutations^{1,6,16–18}. Genome-wide sequencing of plasma samples is used in prenatal diagnostics, demonstrating comprehensive coverage of the genome¹⁹. More recently, genome-wide sequencing of plasma DNA has been

demonstrated as a potential tool for detection of disease or analysis of tumour burden in patients with advanced cancers^{5,7}. These studies established that plasma DNA contains representation of the entire tumour genome⁷, mixing together variants originating from multiple independent tumours⁵. This suggests that deeper sequencing of plasma DNA, applied to selected samples with high tumour burden in blood, may allow assessment of clonal heterogeneity and selection. In this study, we applied exome sequencing of ctDNA as a platform for non-invasive analysis of tumour evolution during systemic cancer treatment (Fig. 1).

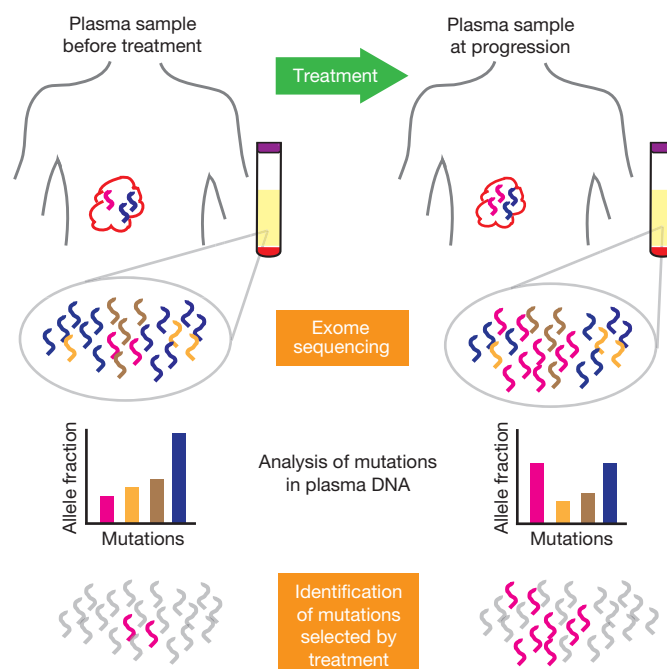


Figure 1 | Identification of treatment-associated mutational changes from exome sequencing of serial plasma samples. Overview of the study design: plasma was collected before treatment and at multiple time-points during treatment and follow-up of advanced cancer patients. Exome sequencing was performed on circulating DNA from plasma at selected time-points, separated by periods of treatment, and germline DNA. Mutations were identified across the plasma samples, and their abundance (allele fraction) at different time-points compared, generating lists of mutations that showed a significant increase in abundance, which may indicate underlying selection pressures associated with specific treatments. These lists contained mutations known to promote tumour growth and drug resistance, but also mutations of unknown significance. Accumulating such data across large cohorts could identify genes or pathways with recurrent mutations.

¹Cancer Research UK Cambridge Institute and University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ²Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. ³Illumina, Inc., Chesterford Research Park, Little Chesterford CB10 1XL, UK. ⁴Department of Haematology-Oncology, National University Cancer Institute, National University Health System, 5 Lower Kent Ridge Road, Tower block level 7, 119074 Singapore. ⁵Cancer Science Institute, National University of Singapore, Centre for Translational Medicine, 14 Medical Drive, #12-01, 117599 Singapore. ⁶Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.

*These authors contributed equally to this work.

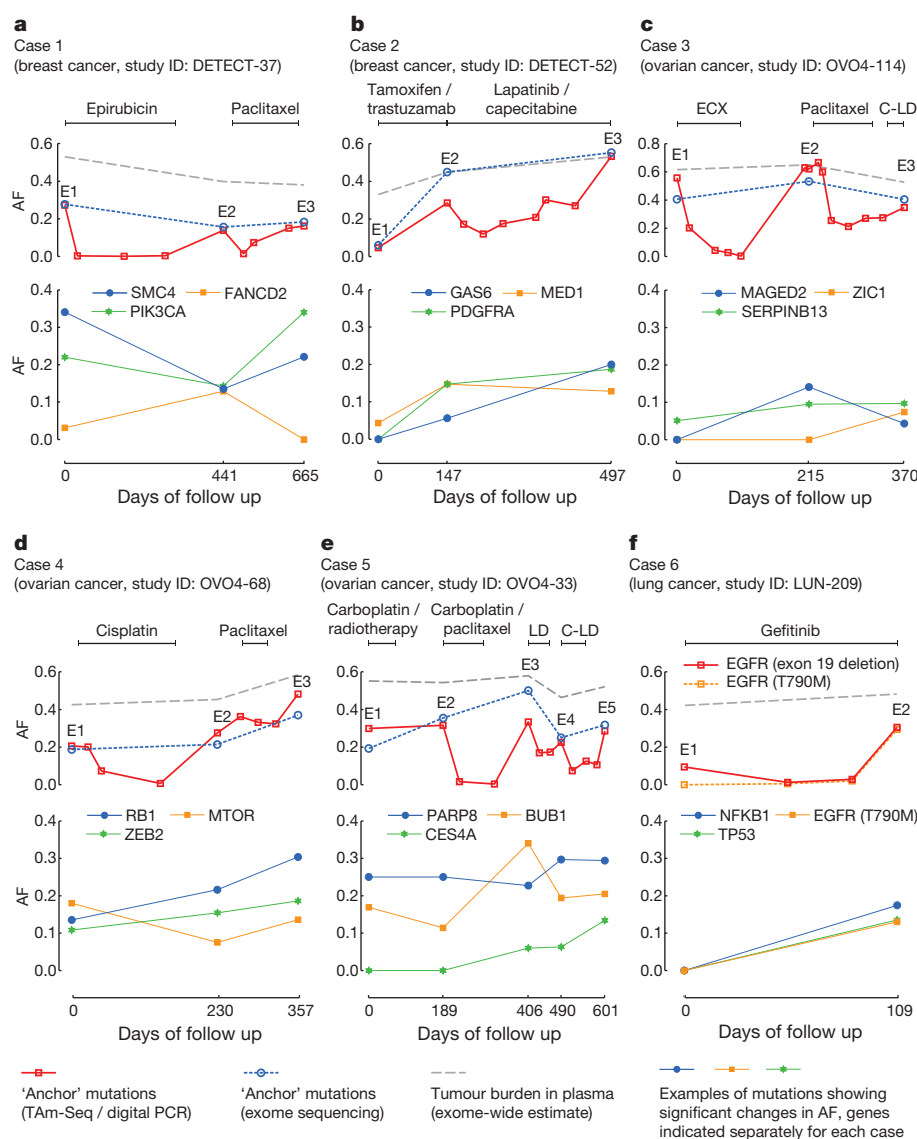


Figure 2 | Mutations showing evidence of genomic tumour evolution. All panels (a–f) are made up of an upper and a lower subpanel. Upper subpanels, time courses for allele fractions (AF; data points) of ‘anchor’ mutations used for initial quantification of ctDNA levels, and the fractional concentration of tumour DNA (tumour burden; grey dashed lines). ‘Anchor’ mutations were measured using digital PCR or TAm-Seq⁶ for all available plasma samples, and using exome sequencing at selected time points indicated by E1, E2, E3 (and E4 and E5 for case 5). Tumour burden was estimated from exome data (an adaptation of genome-wide aggregated allelic loss⁶). In a, AF was averaged over six mutations measured in parallel using digital PCR. In b, a single mutation in

ATM (predicted amino acid change I2948F) was measured by TAm-Seq. In c, d and e, a single mutation in *TP53* was measured by digital PCR for each case (R175H, K132N and R175H, respectively). In f, digital PCR was used to measure abundance of a deletion in exon 19 of *EGFR* (not quantified in exome sequencing data) and the *EGFR* T790M mutation. Lower subpanels, AF in exome data for selected mutations (blue, green and orange datapoints, see key) for each of the cases. Additional details are listed in Table 1, and a full list of mutations that showed a significant increase in abundance is included in Supplementary Tables 2–7. ECX, epirubicin, cisplatin and capecitabine; C-LD, carboplatin and liposomal doxorubicin; LD, liposomal doxorubicin.

We performed whole exome sequencing of plasma DNA in six patients with advanced cancers (Supplementary Table 1): two with breast cancer (cases 1 and 2), three with ovarian cancer (cases 3–5), and one with non-small-cell lung cancer (NSCLC, case 6). Exome sequencing was performed on multiple plasma samples from each patient separated by consecutive lines of therapy, spanning up to 665 days of clinical follow up (range 109–665 days, median 433 days). The ability to detect genomic events using redundant sequencing is dependent on the allele fraction (AF) of the mutant alleles in the samples analysed (ratio of mutant reads to depth of coverage at that locus), the sequencing depth, and the background noise rates of sequencing. Levels of ctDNA were previously quantified in these patients using digital PCR and tagged-amplicon deep sequencing⁶ (TAm-Seq; Fig. 2, upper subpanels), allowing us to focus on samples with a high mutant AF in plasma, in which genomic changes related

to the tumour could be identified even at relatively modest depth of sequencing. Comparison of AF measured using exome sequencing, digital PCR and TAm-Seq showed a high degree of concordance (correlation coefficient 0.8, $P < 0.0001$; Supplementary Fig. 1). Using as little as 2.3 ng of DNA (4%–20% of the DNA extracted from 2.0–2.2 ml of plasma), and an average of 169 million reads of sequencing per sample, we analysed the coding exons of all protein-coding genes at an average unique coverage depth ranging from 31-fold to 160-fold across 19 plasma samples (Supplementary Table 2). Consistent with previous reports^{5,7}, we observed copy number aberrations (CNAs, both gains and losses) in plasma samples in all patients across the whole genome (Supplementary Figs 2–7). These were strongly modulated by the fraction of tumour DNA in plasma and were particularly prominent in plasma samples in which mutant AF exceeded 50%.

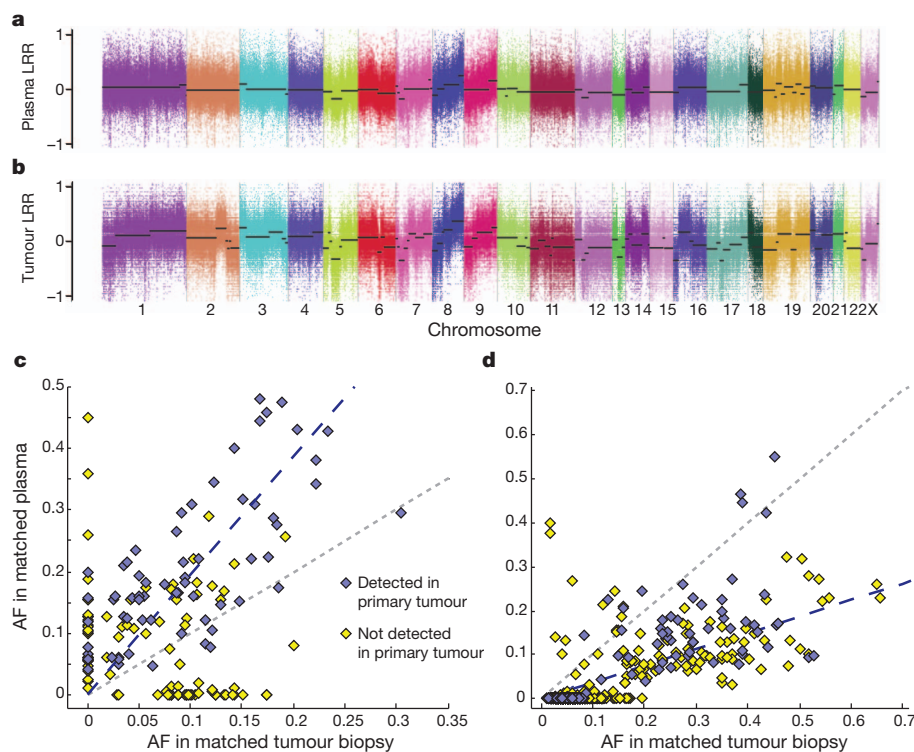


Figure 3 | Genome-wide concordance between plasma DNA and tumour DNA. **a, b,** Sequencing data were used to assess CNAs in the plasma sample (**a**) and in the synchronous metastatic tumour biopsy (**b**) from case 4. Panels show log *R* ratio (LRR), calculated on the basis of exome data, between plasma DNA and normal DNA (**a**) and between tumour and normal DNA (**b**). **c, d,** AF of

mutations identified in exome data from plasma or metastatic biopsy for case 1. Grey dotted line shows equality. Blue dashed line has a slope of 1.93, indicating the median of the AF ratio for mutations found in both samples. Key applies to **c** and **d**. **d,** As **c** but for case 4, blue dashed line has a slope of 0.37.

For two cases, sequencing data were also available from metastatic tumour biopsies, collected at the same time as plasma samples (case 1 sample E1, and case 4 sample E2), and from tumour samples collected at the patients' initial presentation, 9 and 4.5 years earlier. CNAs were concordant between plasma and metastasis DNA in both patients (Fig. 3a, b, and Supplementary Fig. 7). Mutations identified in sequencing data^{20–23} from the plasma or metastatic biopsy were compared (Supplementary Information). In case 1 with breast cancer, 151 mutations were identified in either the plasma or the synchronous biopsy. Of these, 93 mutations were found in both, and mutant AFs for these were higher in the plasma sample compared to the metastatic biopsy. The correlation coefficient of mutant AFs was positive (0.71) for mutations that were also found in the primary tumour, but negative (−0.22) for other mutations (Fig. 3c). In case 4 with ovarian cancer, 895 mutations were identified in either plasma or the tumour biopsy. For 172 mutations found in both, AFs were positively correlated (0.72) and were higher in the metastatic biopsy, which also contained 686 'private' mutations with AF < 0.2 that were not found in either the plasma or the earlier tumour sample (Fig. 3d).

To identify changes in the mutation profiles of the tumours, we compared the abundance of somatic mutations found in plasma before and after each course of systemic treatment. For each patient, we examined a conservative list of mutations, including all mutations that were called in any of the plasma samples with a Bonferroni-corrected binomial probability of <0.05 assuming a background sequencing error rate of 0.1%. For each mutation and course of treatment (spanned by a pair of plasma samples), a *P*-value for a possible change in mutant AF was calculated as the binomial probability of obtaining the observed number of mutant reads, given the sequencing depth and the observed abundance in the paired time-point, normalized by the fractional concentration of tumour-derived DNA in the plasma (based on genome-wide aggregated allelic loss⁵, Supplementary Table 3). Overall, 364 non-synonymous mutations passed with false discovery

rate of <10% for significant changes in normalized abundance, ranging from 15 to 121 for each case (median 49). These include mutations in well-known cancer genes, genes linked to drug resistance and drug metabolism, and genes not previously associated with carcinogenesis or therapy resistance (Supplementary Tables 4–9). Selected examples are shown in Table 1 and Fig. 2.

We highlight here five examples. In case 1 with breast cancer, a strong increase was observed in the abundance of an activating mutation in *PIK3CA* following treatment with paclitaxel (Fig. 2a and Table 1). This mutation has been shown to promote resistance to paclitaxel in mammary epithelial cells⁸. In case 2, a patient with an oestrogen-receptor (ER)-positive, HER2-positive breast cancer, treatment with tamoxifen in combination with trastuzumab led to an increase in abundance of a nonsense mutation near the carboxy terminus of *MED1*, an ER co-activator that has been shown to be involved in tamoxifen resistance^{10,11}. After further treatment of this patient with lapatinib in combination with capecitabine, we observed an increase in abundance of a splicing mutation in *GAS6*, the ligand for the tyrosine kinase receptor AXL (Fig. 2b, Table 1). Activation of the AXL kinase pathway has been shown to cause resistance to tyrosine kinase inhibitors in NSCLC¹³ and resistance to lapatinib in ER-positive, HER2-positive breast cancer cell lines¹². In case 4 with ovarian cancer, following treatment with cisplatin, we observed increase in abundance of a truncating mutation in the tumour-suppressor *RB1* (Fig. 2d, Table 1), predicted to inactivate the RB1 protein (Supplementary Fig. 8). In the matched metastasis biopsy obtained after treatment, the mutation was found in 95% of sequencing reads (59 of 62), with apparent loss of heterozygosity at 13q containing the *RB1* gene (Fig. 3a, b). Loss of *RB1* has been linked with chemotherapy response⁹. Case 6 was a NSCLC patient with an activating mutation in *EGFR* who was treated with gefitinib but progressed on treatment. Analysis by digital PCR detected the *EGFR* T790M mutation in plasma at progression, but not at the start of treatment. This mutation inhibits binding of

Table 1 | Selected mutations whose mutant AF significantly increased following treatment

Patient	Cancer type	Gene	Effect	Potential biological interest	Associated treatment	Mutant AF in plasma	
						Before	After
Case 1	Breast	<i>PIK3CA</i>	E545K	PI-3-kinase. p.E545K mutation associated with chemoresistance in mammary epithelial cells ⁸ .	Paclitaxel	14%	34%
Case 1	Breast	<i>BMI1</i>	S324Y	BMI1 polycomb ring finger oncogene. Associated with chemoresistance ²⁵ .	Paclitaxel	3%	12%
Case 1	Breast	<i>SMC4</i>	I1000S	Structural maintenance of chromosomes 4. Downregulated in taxane resistant cell lines ²⁶ .	Paclitaxel	14%	22%
Case 1	Breast	<i>FANCD2</i>	G56V	Fanconi anaemia complementation group D2. Chromatin dynamics and DNA crosslink repair ²⁷ .	Epirubicin	3%	13%
Case 2	Breast	<i>MED1</i>	S1179X	Mediator complex subunit 1. Co-activator of ER with functional role in tamoxifen resistance ^{10,11} .	Tamoxifen/trastuzumab	4%	15%
Case 2	Breast	<i>ATM</i>	I2948F	Ataxia telangiectasia mutated.	Tamoxifen/trastuzumab	6%	45%
Case 2	Breast	<i>PDGFRA</i>	D714E	Platelet-derived growth factor alpha. Cell surface tyrosine kinase receptor.	Tamoxifen/trastuzumab	0%	15%
Case 2	Breast	<i>GAS6</i>	Splicing	Growth arrest-specific 6. Ligand for AXL, overexpression associated with TKI resistance ^{12,13} .	Lapatinib/capecitabine	6%	30%
Case 2	Breast	<i>TP63</i>	Splicing / S551G	Tumour protein p63.	Lapatinib/capecitabine	4%	20%
Case 4	Ovarian	<i>RB1</i>	E580X	Retinoblastoma 1. Loss of RB1 associated with EMT and drug resistance ⁹ .	Cisplatin	14%	22%
Case 4	Ovarian	<i>ZEB2</i>	Y663C	Zinc finger E-box binding homeobox 2. Overexpression associated with cisplatin resistance in ovarian cancer ²⁸ .	Cisplatin	11%	15%
Case 4	Ovarian	<i>MTOR</i>	K1655N	Mechanistic target of rapamycin. Activating mutations in mTOR confers resistance to antimicrotubule agents ²⁹ .	Paclitaxel	8%	14%
Case 5	Ovarian	<i>CES4A</i>	P55S	Carboxylesterase 4A. Hydrolysis or transesterification of various xenobiotics.	Carboplatin/paclitaxel	0%	6%
					Carboplatin/liposomal doxorubicin	6%	13%
Case 5	Ovarian	<i>BUB1</i>	M889K	Mitotic checkpoint serine/threonine-protein kinase.	Carboplatin/paclitaxel	11%	34%
Case 5	Ovarian	<i>PARP8</i>	P81T	Poly [ADP-ribose] polymerase family, member 8.	Liposomal doxorubicin	23%	30%
Case 6	Lung	<i>EGFR</i>	T790M	Epidermal growth factor receptor. Established to cause gefitinib resistance by inhibiting drug binding ¹⁴ .	Gefitinib	0%	13%
Case 6	Lung	<i>TP53</i>	Y163C	Tumour protein p53 ³⁰ .	Gefitinib	0%	14%
Case 6	Lung	<i>NFKB1</i>	G489V	Nuclear factor κ B ³⁰ .	Gefitinib	0%	17%

Potential biological role and associations with drug resistance described in literature are highlighted. The "Effect" column lists predicted change in amino acid sequence.

gefitinib to EGFR and has been established as the main driver of acquired resistance to gefitinib¹⁴. Unbiased analysis of plasma DNA by exome sequencing identified selection for this mutation amongst genomic changes that occurred following therapy (Fig. 2f, Table 1).

In this proof of principle study, we demonstrate that exome analysis of plasma ctDNA represents a novel paradigm for non-invasive characterization of tumour evolution. Our data, together with recent reports^{5,7}, show that CNAs and somatic mutations identified in ctDNA are widely representative of the tumour genome and provide an alternative method of tumour sampling that can overcome limitations of repeated biopsies. Cell-free DNA fragments from multiple lesions in the same individual all mix together in the peripheral blood⁵, therefore ctDNA is likely to contain a wider representation of the genomes from multiple metastatic sites, whereas mutations present in a single biopsy or minor sub-clone may be missed. This strengthens the case for the use of ctDNA as a biomarker for monitoring tumour burden or for the analysis of hotspot mutation regions^{1,6,16,17}, but also indicates that tracking different mutations for assessment of tumour heterogeneity and clonal evolution is now possible. Our data identified a subset of genes that were positively selected following treatment, many of which have been previously associated with drug resistance. Other changes may represent 'passenger' mutations or false-positives, but some are likely to contribute to resistance to therapy. Accumulating data across a large number of cases could identify new genes or pathways that are frequently mutated following specific treatment types, and help refine analysis algorithms.

The approach we describe here may be broadly applicable to a large fraction of advanced cancers, where the median mutation burden in plasma (before start of treatment) is 5%–10% (refs 6, 16, 24). Analysis of acquired drug resistance is of particular utility in advanced or metastatic cancers, which is the target population for nearly all early phase clinical trials. Improvements in sequencing and associated technologies may enable similar analysis in cases with a lower tumour burden in plasma. At present, this non-invasive approach for characterizing cancer exomes in plasma is readily applicable to patients with high systemic tumour

burden, enabling detailed and comprehensive evaluation of clonal genomic evolution associated with treatment response and resistance.

METHODS SUMMARY

Patients and samples. Cases 1–5 were recruited as part of prospective clinical studies at Addenbrooke's Hospital, Cambridge, UK, approved by the local research ethics committee (REC reference nos 07/Q0106/63, 08/H0306/61 and 07/Q0106/63). Case 6 was recruited as part of the 'Hydroxychloroquine and gefitinib to treat lung cancer' study (NCT00809237) at the National University Health System, Singapore, approved by the National Healthcare Group NHG IRB—DSRB 2008/00196. Written informed consent was obtained from patients, and serial blood samples were collected at intervals of ≥ 3 weeks.

Extraction and sequencing of plasma DNA. DNA was extracted from plasma using the QIAamp circulating nucleic acid kit (Qiagen) according to the manufacturer's instructions. Barcoded sequencing libraries were prepared using a commercially available kit (ThruPLEX-FD, Rubicon Genomics). Pooled libraries were enriched for the exome using hybridization (TruSeq Exome Enrichment Kit, Illumina), quantified using quantitative PCR and pooled in 1:1 ratio for paired-end sequencing on a HiSeq2500 (Illumina).

Variant calling and analysis. Sequencing data were demultiplexed and aligned to the hg19 genome using BWA²⁰. Pileup files for properly paired reads with mapping quality ≥ 60 were generated using samtools²². AFs were calculated for all Q30 bases. A mutation was called if ≥ 4 mutant reads were found in plasma with ≥ 1 read on each strand, and no mutant reads were observed in germline DNA or in a prior plasma sample with ≥ 10 -fold coverage. For comparison between consecutive plasma samples in a patient, we calculated the binomial probability of obtaining the observed AF (or greater) if the abundance of the mutant allele, normalized by tumour load in plasma (based on a modified genome-wide aggregated allelic loss method⁵), had remained constant between the two samples.

Full Methods and any associated references are available in the online version of the paper.

Received 5 October 2012; accepted 11 March 2013.

Published online 7 April 2013.

1. Diaz, L. A. Jr *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).

2. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368**, 842–851 (2013).
3. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
4. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
5. Chan, K. C. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin. Chem.* **59**, 211–224 (2013).
6. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra168 (2012).
7. Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4**, 162ra154 (2012).
8. Isakoff, S. J. *et al.* Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. *Cancer Res.* **65**, 10992–11000 (2005).
9. Knudsen, E. S. & Knudsen, K. E. Tailoring to RB: tumour suppressor status and therapeutic response. *Nature Rev. Cancer* **8**, 714–724 (2008).
10. Cui, J. *et al.* Cross-talk between HER2 and MED1 regulates tamoxifen resistance of human breast cancer cells. *Cancer Res.* **72**, 5625–5634 (2012).
11. Nagalingam, A. *et al.* Med1 plays a critical role in the development of tamoxifen resistance. *Carcinogenesis* **33**, 918–930 (2012).
12. Liu, L. *et al.* Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.* **69**, 6871–6878 (2009).
13. Zhang, Z. *et al.* Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genet.* **44**, 852–860 (2012).
14. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2**, e73 (2005).
15. Punnoose, E. A. *et al.* Evaluation of circulating tumor cells and circulating tumor DNA in non-small cell lung cancer: association with clinical endpoints in a phase II clinical trial of pertuzumab and erlotinib. *Clin. Cancer Res.* **18**, 2391–2401 (2012).
16. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nature Med.* **14**, 985–990 (2008).
17. McBride, D. J. *et al.* Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer* **49**, 1062–1069 (2010).
18. Yung, T. K. *et al.* Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. *Clin. Cancer Res.* **15**, 2076–2084 (2009).
19. Lo, Y. M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
21. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
22. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
23. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
24. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl Acad. Sci. USA* **102**, 16368–16373 (2005).
25. Siddique, H. R. & Saleem, M. Role of BMI1, a stem cell factor, in cancer recurrence and chemoresistance: preclinical and clinical evidences. *Stem Cells* **30**, 372–378 (2012).
26. Chang, H. *et al.* Identification of genes associated with chemosensitivity to SAHA/taxane combination treatment in taxane-resistant breast cancer cells. *Breast Cancer Res. Treat.* **125**, 55–63 (2011).
27. Sato, K. *et al.* Histone chaperone activity of Fanconi anemia proteins, FANCD2 and FANCI, is required for DNA crosslink repair. *EMBO J.* **31**, 3524–3536 (2012).
28. Haslehurst, A. M. *et al.* EMT transcription factors snail and slug directly contribute to cisplatin resistance in ovarian cancer. *BMC Cancer* **12**, 91 (2012).
29. VanderWeele, D. J., Zhou, R. & Rudin, C. M. Akt up-regulation increases resistance to microtubule-directed chemotherapeutic agents through mammalian target of rapamycin. *Mol. Cancer Ther.* **3**, 1605–1613 (2004).
30. Wu, C. C., Yu, C. T., Chang, G. C., Lai, J. M. & Hsu, S. L. Aurora-A promotes gefitinib resistance via a NF- κ B signaling pathway in p53 knockdown lung cancer cells. *Biochem. Biophys. Res. Commun.* **405**, 168–172 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Langmore and K. Solomon (Rubicon Genomics) for early access to library preparation products. We thank L. Jones, S. Richardson, C. Hodgkin and H. Biggs for recruiting patients into the DETECT and CTCR-OVO4 studies, all medical and ancillary staff in the breast and gynaecological cancer clinic and patients for consenting to participate. We thank the Human Research Tissue Bank at Addenbrooke's Hospital which is supported by the NIHR Cambridge Biomedical Research Centre. We thank the Cancer Science Institute, National University of Singapore, and the Hematology-Oncology Research Group, National University Health System, Singapore for their support. We acknowledge the support of Cancer Research UK, the University of Cambridge, National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge Experimental Cancer Medicine Centre, Hutchison Whampoa Limited, and the National Medical Research Council, Singapore. S.-J.D. is supported by an Australian NHMRC/RG Menzies Early Career Fellowship that is administered through the Peter MacCallum Cancer Centre, Victoria, Australia.

Author Contributions M.M., S.-J.D., T.F., D.W.Y.T., D.G., J.D.B., C.C. and N.R. designed the study. M.M., D.W.Y.T. and T.F. developed methods. S.-J.D., C.P., A.S.C.W., T.M.C., J.D.B. and C.C. designed and conducted the prospective clinical studies. M.M., S.-J.D., D.W.Y.T., D.G., T.F. and A.M.P. generated data. Z.K., S.H. and D.B. contributed sequencing data. M.M., F.M. and N.R. analysed sequencing data. S.-F.C. and J.H. contributed to experiments and data analysis. M.M., S.-J.D., D.W.Y.T., T.M.C., J.D.B., C.C. and N.R. interpreted data. M.M. and N.R. wrote the paper with assistance from S.-J.D., D.W.Y.T., C.C., J.D.B. and other authors. All authors approved the final manuscript. J.D.B., C.C. and N.R. are the project co-leaders and joint senior authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D.B. (james.brenton@cruk.cam.ac.uk), C.C. (carlos.caldas@cruk.cam.ac.uk) or N.R. (nitzan.rosenfeld@cruk.cam.ac.uk).

METHODS

Sample collection. Cases 1–5: patients were recruited as part of prospective clinical studies at Addenbrooke's Hospital, Cambridge, UK, approved by local research ethics committee (REC reference nos 07/Q0106/63, 08/H0306/61 and 07/Q0106/63). Written informed consent was obtained from the patients. Serial blood samples were collected in EDTA tubes at intervals of ≥ 3 weeks, and centrifuged within 1 h at 820g for 10 min to separate the plasma from the peripheral blood cells. The plasma was then further centrifuged at 20,000g for 10 min to pellet any remaining cells. The plasma was then stored at -80°C until DNA extraction.

Case 6: this patient was recruited as part of the 'Hydroxychloroquine and gefitinib to treat lung cancer' study (NCT00809237) at the National University Health System, Singapore, approved by the National Healthcare Group NHG IRB-DSRB 2008/00196. Blood was collected in CPT tubes (BD Vacutainer) before gefitinib was started, and at monthly intervals while the patient was on treatment, until disease progression. Blood collected was spun within 1 h at 1,500g for 20 min, and the plasma fraction was frozen at -80°C . Thawed samples were recentrifuged at 20,000g for 10 min to further separate any cellular portions.

Extraction of plasma DNA. DNA was extracted from aliquots of plasma using the QIAamp circulating nucleic acid kit (Qiagen) according to the manufacturer's instructions (see Supplementary Table 1 for volumes used). DNA was eluted into buffer AVE, eluted twice through each column to maximize yield, and stored at -20°C .

Extraction of normal and tumour DNA. DNA from tumour sections was extracted using DNeasy tissue or DNA Allprep kits (Qiagen) according to manufacturer's instructions. Matched germline DNA was derived from normal peripheral blood leucocytes. After the collection of plasma from each blood sample, the remaining layer of normal peripheral blood lymphocytes ('buffy coat') was removed. This layer was either subjected to red cell lysis using a red cell lysis buffer (155 mM NH_4Cl , 10 mM KHCO_3 and 0.1 mM EDTA pH 7.4) and DNA extracted using a standard phenol-chloroform extraction protocol; or frozen at -80°C before extraction using QIAamp DNA mini kit (Qiagen).

Sequencing of plasma DNA. Concentration of DNA for each plasma sample was determined using digital PCR, with an assay targeting *RPP30* for case 2, *TP53* for cases 3–5 and *EGFR* for case 6. For case 1, DNA concentration and 'anchor' mutation AF were calculated by averaging results from six assays targeting *PIK3CA*, *MET*, *IQCA1*, *CD1A*, *KIAA0406* and *ZFYVE21*. Libraries were generated using a commercially available kit for fragmented DNA (ThruPLEX-FD, Rubicon Genomics). 2.3–40 ng of DNA (Supplementary Table 2) was used to generate a sequencing library using manufacturer's protocols. Separate unique molecular identifiers were used for each sample. 30 μl of the library volume was obtained for each sample. 2–5 plasma DNA libraries from each patient were made and pooled together for exome capture using hybridization (TruSeq Exome Enrichment Kit, Illumina). Pools were concentrated using vacuum (Eppendorf Vacuum Concentrator) and prepared to 40 μl volume. Exome enrichment was performed following manufacturer's protocols. Enriched libraries were quantified using quantitative PCR and pooled in 1:1 ratio for paired-end next generation sequencing on HiSeq2500 (Illumina).

Sequencing of normal and tumour DNA. Sequence data for tumour and germline samples for case 1 have been reported previously. In brief, genomic libraries from tumour and matched normal tissue were prepared using the standard Illumina paired-end sample preparation kit according to the manufacturer's instructions. DNA fragments of 300 bp in size were sequenced using paired-end 100 bp reads on a HiSeq2000 (Illumina) achieving a depth of $>30\times$. Germline samples for cases 2–6 and tumour sample for case 4 were sheared using Covaris and exome sequenced as described above.

Digital PCR. The principle of microfluidic digital PCR and its use for quantification of tumour DNA has been described previously^{6,18}. Assays were designed based on TaqMan chemistry. All digital PCR analysis was carried out on the BioMark system using 12,765 Digital Arrays (Fluidigm) following manufacturer's instructions and protocol. Briefly, 3.5 μl from the eluted DNA was heated to 95°C for 1 min and placed on ice, then mixed with TaqMan Universal PCR Master Mix (Applied Biosystems) and sample loading buffer (Fluidigm) into a final reaction volume of 10 μl and loaded into each panel of the chip. The reaction mix was then automatically partitioned into 765 reaction chambers. The numbers of starting template DNA molecules were calculated using Poisson statistics based on the number of positive amplifications^{6,18}.

Analysis of sequencing data. Sequencing reads were demultiplexed allowing zero mismatches in barcodes. Paired-end alignment to the hg19 genome was performed using BWA version 0.5.9 for all exome sequencing data including germline samples, plasma samples and tumour metastasis where generated²⁰. PCR duplicates were marked using Picard. Local realignment was performed using Genome Analysis Tool Kit (GATK)²¹. Pileup files were generated for the genomic regions targeted by exome enrichment using samtools v0.1.17²². For plasma samples, properly paired reads with mapping quality ≥ 60 were used to generate the pileup. AFs for each single-base locus were calculated for all bases with phred quality ≥ 30 .

For germline DNA, an additional pileup file was generated (using a mapping quality cut-off of ≥ 1 and without any base quality cut-offs) and was used as reference for calling somatic variants. A mutation was called if no mutant reads for an allele were observed in germline DNA at a locus that was covered at least 10 fold, and if at least 4 reads supporting the mutant were found in the plasma data with at least 1 read on each strand (forward and reverse). At loci with <10 -fold coverage in normal DNA and no mutant reads, mutations were called in plasma if a prior plasma sample showed no evidence of a mutation and was covered adequately (10 fold or more). All mutations were annotated for genes and function as well as repeated genomic regions using ANNOVAR²³.

AF was defined as the number of high quality reads supporting a mutation as a fraction of the total number of high quality reads covering the locus. For each patient, AF and number of reads for any mutations called with the above parameters were identified in all plasma samples. A binomial probability of obtaining the observed number of reads given depth in each plasma sample was calculated. The minimum of these probability values was corrected using Bonferroni correction for 62 million $\times n$ hypotheses tested, where n was the number of plasma samples sequenced (3 samples for cases 1–4, 5 samples for case 5 and 2 samples for case 6). Mutations with corrected P -values under 0.05 were retained for further analysis in plasma samples.

Estimation of CNAs. To assess CNAs, plasma DNA and tumour sequencing data were compared to germline DNA data at single nucleotide polymorphisms (SNPs) covered within the targeted exome region. The SNPs were identified from the publicly available 1000 Genomes Project data.

Depth information was normalized by dividing the depth of each SNP by the median depth across all SNPs. The log R ratio (LRR) was computed as the base-10 logarithm of the sample depth (metastasis or plasma) divided by the depth of the normal. Each chromosome was segmented by an iterative process that considered non-overlapping blocks of 1,000 data points. Points lying at least 1.5 standard deviations away from the median LRR for the block were removed from the mean LRR computation. If the difference in mean LRR between two consecutive blocks was less than 0.12, the blocks were merged into a single segment whose mean LRR was re-computed using points from both blocks.

Segmentation of B allele frequency (BAF) plots was similarly performed, considering windows of 1,000 data points and starting new segments if the difference in median frequency was greater than 4%. Blocks whose median frequency was within 8% of the median chromosome frequency in the normal sample were considered consistent with the BAF of the normal sample.

Comparison of mutations between plasma and tumour. For tumour/plasma comparison presented for cases 1 and 4, we identified all mutations called in data from synchronous plasma and metastatic tumour samples, as described above. We retained all mutations adequately covered in both samples (minimum 50 reads in plasma, minimum 10 reads in synchronous tumour whole genome data for case 1, minimum 50 reads in synchronous tumour exome data for case 4). We further discarded all mutations with no coverage in archived tumour samples obtained earlier (9 years earlier for case 1, and 4.5 years earlier for case 4).

Identification of mutations that changed in representation over treatment. To estimate systemic tumour burden, we calculated fractional concentration of ctDNA in blood using an adaptation of genome-wide aggregated allelic loss⁵. AFs of SNPs from the 1000 Genomes Project were obtained for germline and plasma data. SNPs with $0 < \text{AF} < 1$ in germline DNA were identified. SNPs where the minor AF in the germline data deviated from heterozygosity were identified using a binomial probability of obtaining the observed number of minor allele reads given depth in germline DNA and expected AF of 0.5. SNPs with probability < 0.25 were discarded from further analysis.

Of the remaining SNPs, significant deviation from heterozygosity in any of the sequenced plasma samples, determined by a binomial distribution using sequencing depth and expected AF of 0.5, was used to identify loss of heterozygosity (LOH). SNPs with a probability < 0.01 in any of the sequenced plasma samples were retained for estimation of tumour burden as described previously⁵. Fractional ctDNA burden was calculated as follows:

$$1 - \left[\frac{\text{sum of reads in the lost alleles}}{\text{sum of reads in the retained alleles}} \right]$$

AFs for all mutations were normalized by the estimated tumour burden. For any comparison between two consecutive plasma samples in a patient, we calculated the binomial probability for the observed difference in AF assuming no difference in normalized abundance. For a comparison between (for example) E1 and E2, we calculated the probability of obtaining the observed number of mutant reads or greater in E2 if normalized abundance in E2 had remained the same as in E1; this probability was multiplied by the probability of the observed number of mutant reads or less in E1 if the normalized abundance in E1 was the same as observed in E2. Where no mutant reads were obtained in the E1, only the reverse direction was used for this analysis. Changes in representation with a false discovery rate of 10% or lower, which were exonic non-synonymous or splicing mutations, were retained and are presented in Supplementary Tables 2–7.

The role of high-throughput technologies in clinical cancer genomics

Expert Rev. Mol. Diagn. 13(2), 167–181 (2013)

Saad F Idris¹,
Saif S Ahmad¹,
Michael A Scott¹,
George S Vassiliou²
and James Hadfield*³

¹Department of Hematology/Oncology,
Cambridge University NHS Hospitals
Foundation Trust, Cambridge,
CB2 0QQ, UK

²Wellcome Trust Sanger Institute,
Hinxton, Cambridge, CB10 1SA, UK

³Cancer Research UK, Cambridge
Research Institute, Li Ka Shing Centre,
Robinson Way, Cambridge, UK

*Author for correspondence:
james.hadfield@cancer.org.uk

Cancer is a genetic disease driven by both heritable and somatic alterations in DNA, which underpin not only oncogenesis but also progression and eventual metastasis. The major impetus for elucidating the nature and function of somatic mutations in cancer genomes is the potential for the development of effective targeted anticancer therapies. Over the last decade, high-throughput technologies have allowed us unprecedented access to a host of cancer genomes, leading to an influx of new information about their pathobiology. The challenge now is to integrate such emerging information into clinical practice to achieve tangible benefits for cancer patients. This review examines the roles array-based comparative genomic hybridization and next-generation sequencing are playing in furthering our understanding of both hematological and solid-organ tumors. Furthermore, the authors discuss the current challenges in translating the role of these technologies from bench to bedside.

KEYWORDS: array CGH • cancer • cancer genomics • next-generation sequencing • oncology
• personalized medicine • pyrosequencing • SNP-CGH • targeted therapy

Genomic alterations in cancer

A small number of hereditary cancer syndromes are directly caused by germline mutations and heritable genetic variation may also play a role in many sporadic cancers. However, the great majority of human cancers are driven by somatic mutations within the cancer-cell genome that occur during life. Understanding the diversity and function of these somatic mutations is the cornerstone of current cancer research and detecting known mutations of clinical significance reliably is increasingly forming an important part of clinical practice. The major impetus for elucidating the nature and function of somatic mutations in cancer genomes is the potential for the development of effective targeted anticancer therapies, the archetypal example being the tyrosine kinase inhibitor imatinib, which directly inhibits the *BCR-ABL* fusion gene product arising from the translocation t(9;22) in chronic myeloid leukemia, and which has revolutionized the treatment and outcome of this previously devastating disease [1].

Nucleotide substitutions are the most common genomic alterations in tumors – usually at a stated rate of one substitution per million nucleotides [2]. Insertion and deletions are

ten-times less common. The rate of mutation varies significantly. For example, skin melanomas occurring as a result of UV radiation exposure, display substantially more mutations than hematopoietic tumors [3–5]. Even small point mutations or microdeletions can be vitally important to detect, as they may have major relevance to the patient's prognosis and future treatment [6]. Much larger acquired chromosomal translocations are a well-characterized feature of hematological malignancy but have also been demonstrated in solid-organ tumors [7,8]. Smaller copy-number variations (CNVs) in tumor genomes can also result in the amplification of oncogenes and/or inactivation of tumor-suppressor genes contributing directly to tumor pathogenesis. Additionally, loss of heterozygosity (LOH) of tumor suppressor genes is increasingly being recognized as an important genomic alteration contributing to tumor initiation and progression [9]. After an inactivating mutation of the first allele, LOH occurs either as a result of loss of function of the remaining normal allele or as a result of uniparental disomy where the mutant allele is duplicated and the remaining normal allele lost. Reliable detection of all of these types of mutation within cancer genomes of individual

patients is necessary if we are to offer truly personalized cancer care to our patients.

Current methods

The mainstay of current cancer diagnosis is histological examination of tumor cells with immunohistochemical analysis. These methods are relatively blunt instruments that fail to distinguish between molecularly distinct subtypes of tumors, which may have individual tumor biology, prognosis and treatment. In a variety of cancers, especially hematological, more specific molecular tests are now widely employed to provide more detailed information about individual patients' disease. It is routine practice, for example, to examine bone marrow samples from patients with acute leukemia using flow cytometry, in order to identify the pattern of cell-surface-marker expression characteristic of the different subtypes. Furthermore, metaphase cytogenetics and FISH can be used to detect gross chromosomal abnormalities, which can then be used to risk-stratify patients and, in many cases, to guide their treatment. Previously, no systematic approach had been adopted to study complex karyotypes in solid-organ tumors; however, recent discoveries of important translocations in a variety of tumors have highlighted the importance of detecting such aberrations [7,10].

Traditional sequencing techniques remain the mainstay of detecting germline mutations responsible for hereditary forms of cancer, for example, *BRCA1* and *BRCA2* mutations in breast cancer. More recently, DNA sequencing has expanded into routine clinical practice to assess tumor cells for specific mutations, particularly with respect to their response to various targeted therapies (TABLE 1). To date, most diagnostic laboratories use automated versions of the classical 'chain termination' method described by Fred Sanger in 1977 to determine DNA sequence. Pyrosequencing is a common alternative method that relies on detection of a chemiluminescence signal released by a luciferase enzyme as a pyrophosphate group is released by DNA polymerase at the addition of each nucleotide. Pyrosequencing is more sensitive than the Sanger method but can only sequence shorter DNA templates. Hence, it is best used clinically for hotspot sequencing of short

DNA/exon sequences where known mutations are commonly found. Pyrosequencing is also useful for use with formalin-fixed, paraffin-embedded (FFPE) tissue sections, which usually yield short fragmented DNA. Overall, owing to the limited bandwidth and throughput of such first-generation techniques, they have only been applied clinically in a targeted way to look for small numbers of known mutations with established clinical relevance.

However, with growing knowledge of many different genes that contribute to the pathophysiology of a particular tumor, there has been a shift in focus to genome-wide techniques that can interrogate a larger proportion of the cancer genome in a more unbiased way. A range of technologies can be used in this way, including gene expression profiling, array-based comparative genomic hybridization (aCGH), SNP-CGH and next-generation sequencing (NGS). These techniques are well established in cancer research and have offered remarkable insights into tumor biology. There is thus a growing expectation that these technologies should now become incorporated into clinical practice and bring about a real change to the diagnosis and treatment of individual cancer patients' management. In the authors' opinion, CGH (including both oligonucleotide array CGH, aCGH and SNP probe-based SNP-CGH) and NGS are the two high-throughput technologies at the forefront of making this transition (FIGURE 1). They both possess particular strengths and weaknesses, which will influence their clinical utility in the future.

aCGH/SNP-CGH

CGH was developed for molecular cytogenetic analysis of solid tumors, and it has developed significantly over the past two decades, moving to array-based formats [11]. The first of these were bacterial artificial chromosome (BAC) arrays; however, these have limited chromosomal resolution and have proven difficult to manufacture on a commercial scale [12–14]. Because of this, BAC arrays have been replaced almost entirely by arrays of oligonucleotides, which can be manufactured more reproducibly and at very high probe densities, allowing almost base-pair resolution. Oligonucleotide array CGH (aCGH) is normally performed

Table 1. List of US FDA approved cancer drug therapies.

Disease	Mutation	Drug therapy	Current test platform
DLBCL	<i>c-myc</i>	R-CODOX-M-IVAC chemotherapy	PCR
CML	<i>BCR-ABL</i> translocation	Imatinib/dasatinib/nilotinib	Cytogenetics/FISH/PCR
AML (promyelocytic)	<i>RARA-PML</i> mutation	All- <i>trans</i> retinoic acid	Cytogenetics/FISH/PCR
Breast cancer	<i>HER2</i> amplification	Trastuzumab/lapatinib/pertuzumab	IHC/FISH
NSCLC	<i>EML-ALK</i> translocation; <i>EGFR</i> mutation	Crizotinib; gefitinib/erlotinib	FISH; traditional sequencing
Colorectal cancer	<i>KRAS</i>	Cetuximab/panitumumab	Traditional sequencing
Gastric cancer	<i>HER2</i> amplification	Trastuzumab	IHC/FISH
Melanoma	<i>BRAF</i> mutation	Vemurafenib	Traditional sequencing
Medullary thyroid cancer	<i>RET</i> mutation	Vandetanib	PCR
GIST	<i>c-kit</i> mutation	Imatinib/sunitinib	IHC

AML: Acute myeloid leukemia; CML: Chronic myeloid leukemia; DLBCL: Diffuse large B-cell lymphoma; GIST: Gastrointestinal stromal tumor; IHC: Immunohistochemistry; NSCLC: Non-small-cell lung cancer.

Adapted with permission from [77].

using either arrays of long-oligos designed to hybridize with specific genomic loci (aCGH) or using arrays of oligos designed to report specific SNP genotypes (more commonly referred to as SNP-CGH and developed by companies such as Affymetrix [CA, USA] and Illumina [CA, USA]). This can be confusing as both technologies require a comparison to be made to generate data for cancer diagnostic or prognostic use. The authors use the terms aCGH and SNP-CGH to point readers to the differences in the technologies.

aCGH allows detection of copy number differences between a test and reference sample of DNA. Both samples are labeled with different fluorophores and then added to a glass slide containing several thousand (up to and over 1M) fixed oligonucleotide probes dispersed evenly throughout the genome (they can also be targeted to be denser in known regions). The samples hybridize to the probes and the relative fluorescence intensity from the test and reference sample are compared in order to ascertain the CNVs at a particular locus. Oligonucleotide probes are particularly advantageous, as they can be standardized across all arrays used, are devoid of repetitive sequences, and are subsequently much more reproducible. They can be spaced more densely across specific parts of the genome, allowing for better detection of smaller genomic changes and providing increased sensitivity. They can also be customized as information about the genome is updated, and multiple probes can target a single region, allowing for more robust data analysis and increasing reproducibility, sensitivity and confidence in CNV calls [15]. Using this method, copy-number changes affecting regions as small as 5–10 kb can be detected. High-resolution CGH arrays are now available that allow accurate detection of structural variations at resolutions of 40–80 bp, appropriate for detection of microdeletions and duplications [16]. However, small sequence alterations or single basepair mutations will still not be detected; neither will balance chromosomal translocations or inversions for which FISH remains an important technique [17]. Another disadvantage of standard aCGH is its relative inability to detect areas of LOH. These can be detected in cancer cells by noticing the presence of heterozygosity at a particular genetic locus in the germline DNA but the absence of it at the same locus in the cancer-cell genome. This can be more readily ascertained by using SNP-CGH analysis designed for genome-wide association studies (GWAS) to form a virtual karyotype. SNP-CGH is a related microarray technology that uses oligonucleotide probes corresponding to allelic variants of selected SNPs [18]. Hybridization of genomic DNA to both probe

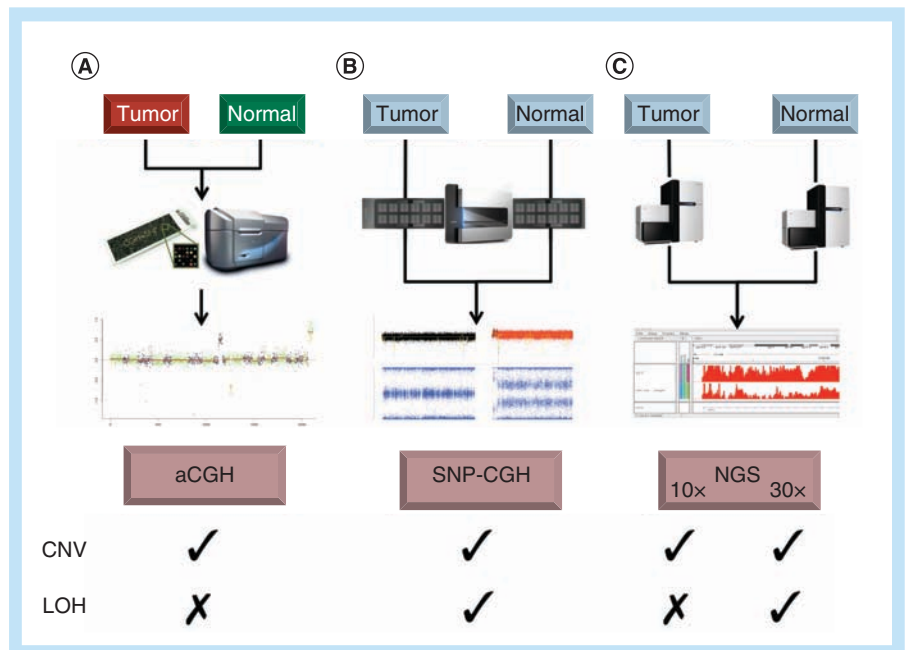


Figure 1. Workflows and suitability of genomics tools for copy-number variation and loss-of-heterozygosity. (A) Two-color aCGH microarrays: tumor and normal DNAs are differentially labeled and applied to the same slide for direct estimation of CNV from the resulting probe signals, but not LOH. (B) One-color SNP-CGH microarrays: tumor and normal DNAs are labeled and applied to separate arrays; *in silico* analysis allows both CNV and LOH analysis to be completed. (C) NGS: tumor and normal DNAs are converted to sequencing libraries, which are sequenced to a specified depth or coverage (the average number of times each base is read); *in silico* analysis allows CNV calls to be made for 10× coverage data and both CNV and LOH from 30× coverage data. aCGH: Array-based comparative genomic hybridization; CGH: Comparative genomic hybridization; CNV: Copy-number variation; LOH: Loss of heterozygosity; NGS: Next-generation sequencing.

variants indicates heterozygosity, while a signal for only one allele indicates either homozygosity or LOH. This technique can also use the intensity of the genotype signal and B-allele frequency to determine DNA copy number. Although individual SNP probes have significantly lower sensitivity to detect CNV than BAC or oligonucleotide probes, they are generally present in much higher numbers, thus compensating for their lower sensitivity. SNP arrays commonly make use of 1 million probes or more, allowing data analysis to determine copy-number status over multiple probes [19]. Both aCGH and SNP-CGH arrays are analyzed using multiple probes to generate copy-number data for genomic loci. The circular binary segmentation algorithm was developed for copy-number analysis using microarrays [20]. The resolution of the array is determined by the type and number of probes present on the array, and on the analysis method used (TABLE 2). The same basic algorithm has been developed for use with NGS data [21].

NGS

Sanger sequencing was used for the Human Genome Project [22], but despite the significant technical improvements to this 'first-generation' technology, new second-generation screening or NGS technologies are required for sequencing multiple human genomes at adequate depth. Over the last 5 years, three companies have

Table 2. Commercial array-based comparative genomic hybridization and single nucleotide polymorphism-comparative genomic hybridization products.

Company	Platform	Array/ sequencing technology	Median probe spacing (kb)	Probes per call (n)	Effective resolution (kb)	Probes (n)	Oligonucleotide length (mer)	Detection	Custom content	DNA input	TAT (days)	HOT
Agilent	aCGH	4 x 44K	43	5	215	43,000	60	CNV	Yes	0.1 lg	3	
	aCGH	244K	9	5	45	236,000	60	CNV	Yes	0.5–1.5 lg	3	
	aCGH	2 x 400K	5.5	5	27.5	411,000	60	CNV	Yes	0.5–1.5 lg	3	
	aCGH	1 million	2	5	10	963,000	60	CNV	Yes	0.5–1.5 lg	3	
	aCGH	Cancer CGH and SNP 4 x 180k [†]	25.3	1–3	40	180,000	60	CNV, genotype, and LOH	Yes	500 ng	3–4	~7 h for 8 samples
Affymetrix	aCGH	Cancer CGH and SNP 2 x 400k [†]	7.2	1–3	15	400,000	60	CNV, genotype, and LOH	Yes	500 ng	2–3	~7 h for 8 samples
	aCGH	ISCA CGH+SNP bundle, 4 x 180k [†]	25.3	1–3	40	180,000	60	CNV, genotype, and LOH	Yes	500 ng	2–3	~7 h for 8 samples
Affymetrix	SNP-CGH	SNP array 6.0 [‡]	0.7	10	7	1,800,000 [‡]	25	CNV, genotype, and LOH	No	500 ng	2–3	~8 h for 96 samples
	SNP-CGH	CytoScan® HD [‡]	0.7	10	7	2,700,000 [‡]	25	CNV, genotype, and LOH	No	250 ng	2–3	~8 h for 96 samples
Illumina	SNP-CGH	CytoSNP-12 [‡]	6.2	10	62	300,000 [‡]	50	CNV, genotype, and LOH	Yes	200 ng	2–3	~8 h for 24 samples
	SNP-CGH	Human660- Quad [‡]	2.3	10	23	660,000 [‡]	50	CNV, genotype, and LOH	Yes	200 ng	2–3	~8 h for 24 samples
	SNP-CGH	Human1M- Duo [‡]	1.5	10	15	1,200,000 [‡]	50	CNV, genotype, and LOH	Yes	200 ng	2–3	~8 h for 24 samples [‡]

[†]Costs include array and reagents.

[‡]Mix of SNP and copy-number variation probes.

aCGH: Array-based comparative genomic hybridization; CGH: Comparative genomic hybridization; CNV: Copy-number variation; HOT: Hands-on time; ISCA: International Standards for Cytogenomic Arrays; LOH: Loss of heterozygosity; TAT: Turnaround time.

provided the most widely used NGS systems: Illumina, Roche (NJ, USA) and Life Technologies (CA, USA). An overview of the various NGS technologies is provided below, but the authors would refer readers to an excellent review by Baylor College of Medicine's Michael Metzker for further details [23].

Roche/454: pyrosequencing

454 Life Sciences Corp. developed the first NGS technology and fundamentally changed perceptions of what might be achieved with sequencing [24], and in 2010, the first NGS human genome was published using 454 sequencing [25]. Libraries are prepared by ligating oligonucleotide adapters to fragmented genomic DNA. DNA is denatured and single-stranded adapter-ligated fragments are hybridized to micrometer-sized beads. DNA fragments are amplified on the beads in an emulsion-PCR, resulting in beads carrying tens of millions of copies of the original DNA fragment. After PCR, the emulsion is broken, and DNA-coated beads are purified, denatured and loaded into the wells of a 'picotiter' plate. The wells of the picotiter plate are large enough for only a single bead to be loaded; each well carrying a bead will generate an individual DNA sequence. Pyrosequencing is performed by cyclical addition of individual nucleotides, sulfurylase and luciferase. As each nucleotide is incorporated into the growing strand, an inorganic pyrophosphate group is released and converted to ATP by the sulfurylase. Luciferase uses the ATP to convert luciferin to oxyluciferin, producing a light signal that is directly proportional to the number of inorganic pyrophosphate molecules released and the number of nucleotides incorporated. The first publication generated 250,000 reads of 80–120 bp in length [17]. Around 1 million sequences of up to 700 bp in length are currently generated in a GS-FLX (Roche) run.

Illumina method: sequencing by synthesis

The Illumina HiSeq is the most widely adopted NGS instrument to date and was used to sequence the first cancer genomes [5,26]. Illumina have significantly refined the sequencing-by-synthesis technology (SBS) it acquired from Solexa in 2006, improving chemistry, instruments and software. Libraries are prepared by ligating Y-shaped oligonucleotide adapters. These are prepared from two oligonucleotides that share complementarity at one end; when annealed and ligated to DNA fragments they allow different sequences to be added to the end of each fragment. The library of DNA fragments is enriched by PCR ready for clustering and sequencing. Libraries are denatured to pM concentration and are introduced to an Illumina flowcell; the fragments hybridize to complementary oligonucleotides on the surface of the flow cell and are copied by DNA polymerase. These daughter molecules are then 'bridge-amplified' by repeated cycles of chemical denaturation and polymerase extension to produce discrete clusters each containing about 1000 molecules. SBS uses fluorescently labeled and reversibly blocked terminator deoxynucleoside-triphosphates in a cyclic sequencing reaction. Nucleotides are incorporated by DNA polymerase into the growing DNA strand, the flow cell is imaged to determine which nucleotide has been incorporated into each individual cluster, and finally the terminator is removed

by chemical cleavage ready for the next round of incorporation, imaging and cleavage [27]. The early Solexa-based sequencers from Illumina generated reads of 35 bp in 2007 and generated around 30 million sequences or 1 Gb of data from a flow cell. Read length has increased to 150 bp on the HiSeq 2500 system, which generates over 1.5 billion sequences (or 3 billion paired-end sequences) and 300 Gb of data from a single flow cell as of December 2012. The cost of sequencing a human genome on the HiSeq 2000 was estimated to be just US\$6500 in 2011 [28]. The authors estimate that at the time of writing the cost of a 30x human genome was US\$4000, and the cost of a 10x human genome was US\$1300. Illumina released a lower throughput personal genome sequencer, the MiSeq in 2011.

Life Technologies: sequencing by ligation

Life Technologies initially developed the Agencourt Personal Genomics support oligonucleotide ligation detection (SOLiD™) sequencing technology in their SOLiD 3, 4 and 5500 instruments, but these have not seen widespread adoption by the sequencing community due to reduced throughput and a more complex workflow. The SOLiD system uses emulsion PCR to generate template beads for sequencing by ligation. Beads are then deposited onto a slide and primers hybridize to the adaptor sequence on the template beads. Four fluorescently labeled probes compete for ligation to the sequencing primer. Multiple cycles of ligation, detection and cleavage are performed, with the number of cycles determining the eventual read length of up to 75 bp. More recently, Life Technologies acquired Ion Torrent to release the PGM™ and Proton™ sequencers. These systems use a very similar approach to the original 454 pyrosequencing. The sequencing is performed on a semiconductor chip that has wells into which individual emulsion PCR beads can be loaded. Sequencing is performed in a similar cyclical manner, but as each nucleotide is incorporated hydrogen ions are released, which change the pH of the well. This is detected by the ion sensors in the semiconductor chip, which then produces a 'flowgram' format similar to the Roche/454 'pyrogram'. Life Technologies effectively obsoleted their own SOLiD technology with their Ion Torrent products.

Personal NGS instruments & targeted resequencing

The development of benchtop 'personal' NGS instruments such as MiSeq (Illumina) and PGM (Ion Torrent) present real opportunities for the use of NGS in a clinical setting. The authors have not included GS Junior (Roche/454) in this list as the authors do not believe it will compete in the long term against the other technologies. These bench-top instruments offer several advantages over the larger 'whole genome' sequencing instruments. They are generally much cheaper to buy, they are very much faster to run and the volume of data generated is smaller and therefore easier to manage. Both the MiSeq and PGM can perform multiple sequencing runs in a day, offering laboratories greater throughput and flexibility. This comes with a reduction in the Gb of sequence data generated. Rather than 100s of Gb, only single digit or tens of Gb are generated. This is not the obvious drawback that it may

seem as the clinically relevant portion of the genome is currently quite small. Although still quite new, the technologies are already being compared; however, this may be premature given the rapid pace of development [29].

Many groups are taking advantage of this by adopting targeted sequencing of specific regions of interest. There are multiple methods to target the genome, the simplest of which is PCR. Methods differ in the amount of target that can be captured, in the amount of DNA used and in the throughput, cost and time of a single assay. Choosing a target-enrichment strategy will depend on project-specific requirements and personal preferences. Many laboratories have been using targeted resequencing, an excellent example of which is the comparison of technologies and the use in a screening test for carriers of 448 severe childhood recessive illnesses [30]. The authors of this study also discussed the need for confirmatory testing and suggest that the high confidence achieved with 10× coverage made this unnecessary. These applications are developing rapidly and the authors discuss them further in the 'Expert commentary' of this review.

Clinical applications of aCGH & NGS

Screening

Screening at-risk populations for cancer-susceptibility genes is an evolving field. Germline mutations in the tumor-suppressor genes *BRCA1* and *BRCA2* are known to predispose breast and ovarian cancers. Less than 1% of the population carry *BRCA* mutations, however, lifetime risks of breast cancer are as high as 80% among affected women [31]. Although recent data have demonstrated that family history is not a reliable indicator of *BRCA* status, women with a strong family history of these cancers are offered genetic testing, and mutation carriers may be offered prophylactic oophorectomy and bilateral mastectomy [32,101]. At present, the majority of *BRCA* testing is performed using the US company Myriad's BRACAnalysis® technique, which uses a combination of PCR and traditional Sanger sequencing [102]. Recently, a number of laboratories have developed NGS-based *BRCA* tests, which can concurrently sequence multiple other-candidate genes with a faster turnaround time and reduced cost compared with BRACAnalysis [33]. Use of these tests is currently limited, however, by Myriad's patenting of the *BRCA1* and *BRCA2* genes.

Diagnosis & prognostication

Microarray technology is increasingly being used to compare tumor and germline DNA from cancer patients, allowing detection of somatic lesions acquired by the tumor cells. A large number of studies have employed both aCGH and SNP arrays as global approaches to detect CNVs and LOH in a research setting. In a clinical context, these techniques are particularly applicable to the diagnosis of hematological malignancies where chromosomal aberrations are well described. For example, a recurrent area of difficulty in clinical practice is the accurate diagnosis of lymphoma subtypes with the current combination of histological, immunohistochemical and targeted molecular techniques. Chromosomal translocations, particularly those involving the immunoglobulin loci, are a hallmark of many types of B-cell

lymphomas, and specific disease entities display characteristic genomic alterations by which they can be distinguished [34–36]. These observations led Takeuchi and colleagues to use aCGH to generate a computational differential diagnosis [36]. With this method, they correctly classified 88% of the diffuse large B-cell lymphomas (DLBCL) and mantle-cell lymphomas and 83% of the activated B-cell and germinal center B-cell subtypes in their cohort. These results demonstrate that CNVs detected by aCGH can be used for classifying lymphomas into biologically and clinically distinct diseases or subtypes. With growing knowledge of disease-specific chromosomal rearrangements, it may be possible to apply aCGH techniques to routine diagnosis of lymphoma subtypes, although this remains to be validated in more substantial clinical studies.

Prognostic markers of disease benefit both patients and doctors, enabling better-informed decisions regarding treatment. Despite previous attempts at subclassification, until recently, DLBCL remained a biologically heterogeneous tumor with no clear prognostic biomarkers [37]. Analysis of 40 patients with DLBCL showed that aCGH not only reliably detected CNVs, but in addition the investigators observed a different cytogenetic profile in those patients achieving complete versus partial remission. Such approaches have the potential of using high-resolution genome scanning to identify new regions associated with poor outcome, and could help with stratification of patients with aggressive lymphoma in the future [38]. Similarly, high-resolution aCGH has been successful in detecting CNVs that are exclusive to either chemoresistant or chemoresponsive DLBCL [39]. More recently, *MYC* gene rearrangements were shown to be of prognostic significance in patients with DLBCL, and although currently detected by PCR-based and FISH techniques, would be reliably detected by aCGH methods [40]. Prognostication based on cytogenetic findings is common practice in many other hematological malignancies, including acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL) and chronic lymphocytic leukemia (CLL). A recent study validated the use of a 250,000-SNP array for clinical use by demonstrating 98.5% concordance with a standard CLL FISH panel. SNP array karyotyping also detected areas of LOH not detected by the FISH panel, which would also have remained undetected with aCGH [41].

aCGH techniques have been applied in a range of other hematological malignancies, albeit from a research/gene-discovery perspective. Targeted diseases include multiple myeloma, AML and myelodysplastic syndromes (MDS) [42–44]. In general, these results have revealed a number of more complex chromosomal defects than detectable with metaphase cytogenetics (MC). Of particular importance were microdeletions involving specific genes; for example, 4q involving *TET2* in MDS. Analysis of 140 patients with AML using MC and SNP array demonstrated a clear advantage of SNP array in detecting unbalanced lesions and areas of uniparental disomy. In addition, this enhanced detection led to improved predictive power even for known mutations such as the *FLT3* internal tandem duplication [45]. Similar results have been reported for MDS [46]. In a cohort of 430 patients with MDS, combined MC/SNP array karyotyping has conclusively shown

a higher diagnostic yield of chromosomal defects (74 vs 44%; $p < 0.0001$), compared with cytogenetics alone, often through detection of novel lesions. The presence and number of new SNP array-detected lesions were in themselves independent predictors of overall and event-free survival [47]. This underscores the significant diagnostic and prognostic contributions of SNP array-detected defects in MDS and related diseases [48].

The evidence for clinical application of NGS techniques is also mounting and is highlighted in a dramatic case where standard MC failed to identify a *PML-RARA* fusion event in a patient with acute promyelocytic leukemia whose absence would have necessitated a more aggressive treatment regimen culminating in allogeneic bone marrow transplantation. With a high index of suspicion, the responsible clinicians arranged for whole-genome NGS and successfully identified a cryptic gene fusion within 7 weeks of biopsy. The patient was thus spared an allogeneic bone marrow transplant with its high attendant treatment-related mortality [49].

With regards to solid tumors, outcomes in cancers of unknown primary origin are most likely to benefit from NGS-based technologies. Cancers of unknown primary origin are diagnosed in cases when a metastatic lesion is found in a patient in whom a primary lesion cannot be identified despite appropriate investigations [50]. It is usually associated with a poor prognosis, in part due to the uncertainty of the diagnosis. At present, a number of gene expression-based arrays are being investigated in clinical trials and it is probable that NGS will play an increasing role in this field.

Treatment, response & relapse

The emergence of small-molecule inhibitors and antibodies against 'druggable' gene targets is revolutionizing cancer therapeutics. The success of these therapies depends on the genetic profile of the individual tumor being treated. Personalizing anti-cancer therapy therefore relies upon identifying each patient's cancer-specific driver mutations.

The significance of *BRCA* mutations, for example, is not limited to screening alone, as chemotherapy agents such as cisplatin and PARP inhibitors have demonstrated greater efficacy in *BRCA*-deficient tumors [51]. The ALK inhibitor crizotinib is highly effective in advanced non-small-cell lung cancers that feature *EML4-ALK* fusion [52]. A growing number of such somatic cancer mutations are being identified. Of particular clinical significance currently are mutations in *KRAS* and *EGFR* genes within the context of advanced colorectal cancer and non-small-cell lung cancer (NSCLC), respectively. Colorectal cancer patients with mutant *KRAS* fail to derive benefit from anti-EGFR therapies (e.g., cetuximab) [53]. Conversely, NSCLC patients with somatic mutations in *EGFR* who receive the EGFR inhibitor, gefitinib, have far superior outcomes to patients receiving standard chemotherapy [6]. Although as yet there is no clear consensus on how best to identify *EGFR* and *KRAS* mutations, in the clinic a number of sequencing-based methods are being used [54,55]. Recently, targeted NGS of 24 NSCLC FFPE tissue specimens identified a *KIF5B-RET* gene fusion in one sample. The fusion gene was seen in 11 out of 561 additionally screened tumors. RET inhibition with multitargeted tyrosine kinase inhibitors represents a

promising treatment in these patients. The study emphasizes that NGS can have significant clinical application, even using minimal tissue from FFPE tumor biopsies [56]. It is inevitable that with time, further such mutations will be identified and while currently single-agent targeted drug therapies are commonly used, future therapies will rely on targeted drug combinations with the aim of improving efficacy and reducing drug resistance, particularly as it is increasingly being recognized that many of the mutations within an individual cancer type are also present to variable degrees in multiple other cancer types [57].

Although not part of routine clinical practice, there are tantalizing glimpses of how sequencing technology may be applied within the clinic. Researchers in Boston (MA, USA) have developed a multiplexed PCR-based assay (SNaPshot™), which identifies more than 50 mutations in a number of important NSCLC genes. Over 15 months, they used the SNaPshot assay to genotype 552 NSCLCs as part of standard care. More than 50% of cases were positive for a driver mutation. In over 30 patients, a less common mutation was identified for which there was a plausible targeted therapy (e.g., *BRAF* and *HER2*). In one case, a patient presented with a contralateral lung lesion 2 years after curative surgery. The second lesion was genetically distinct from the previous primary, making the patient eligible for aggressive therapy. The authors concluded multiplexed genotyping to be a clinically feasible approach to support diagnostic and treatment decisions and to facilitate clinical trial enrollment [58].

Another pertinent example is *BRAF* in malignant melanoma. Metastatic melanoma has a devastating prognosis with median survival from diagnosis of 8–18 months. In 2002, a landmark paper identified mutations in the *BRAF* oncogene in 59% of melanoma cell lines [59]. The majority of mutations were reported in *BRAF* exon 15: T1796A leading to substitution of valine by glutamic acid (V600E). Mutations in *BRAF* cause constitutive activation of downstream signaling through the MAP kinase pathway. Consequently, drugs targeting BRAF were developed and the most promising of these in early-phase clinical trials was vemurafenib (PLX4032). These findings were confirmed in a randomized Phase III trial where response rates in the vemurafenib arm were 48% compared with 5% in the standard chemotherapy arm. Despite high response rates, the duration of response to targeting BRAF in melanoma is disappointingly short-lived at a median of around 6.7 months, indicating acquired resistance [60]. For most targets, acquired resistance usually occurs due to secondary mutations within the target [61]. However, in the case of *BRAF* sequencing, resistant tumors demonstrated no evidence of such mutations [62]. Wagle *et al.* subsequently used targeted resequencing of melanoma from an individual patient to identify a new mechanism of acquired resistance to vemurafenib [63]. Hybridization capture and resequencing of 138 cancer genes in the tumor samples taken before and after relapse revealed a new mutation in the MEK1 kinase only present in the relapse sample. It is extremely unlikely that such novel mutations would have been discovered by traditional genotyping approaches. NGS may thus play a vital role, not only in planning patients' therapy but also in understanding and ultimately bypassing mechanisms of drug resistance. A number

of other resistance mechanisms have now been identified, including upregulation of *N-RAS* or *PDGFR* and methods to overcome these in the clinical setting are being examined [62]. Such studies are leading to a paradigm shift in the management of melanoma specifically, and in cancer trial design as a whole.

Radiotherapy (RT) has an important role in anticancer therapy and also stands to benefit from advances in gene technology. RT is involved in the management of 40% of cancer patients cured of their disease. However, toxicity within healthy tissue remains one of its main limitations. In 2009, a Radiogenomics Consortium was established within the UK with the main aim of identifying genetic variants, primarily SNPs, associated with the development of normal tissue toxicities resulting from radiation therapy [64]. Radiogenomics is a field in its infancy, but already the idea of tailoring RT doses based on radiosensitivity of the primary tumor and adjacent healthy tissue is beckoning. The first GWAS in a large group of RT patients has demonstrated no associations with late toxicity for any of the candidate SNPs, emphasizing that further research in this area is much needed [65].

Although routine analysis of cancer genomes using NGS is an aim for the short to medium term, more imminently clinical trials are likely to incorporate high-throughput sequencing to generate comprehensive, individual mutational landscapes. A recent pilot study explored the practical challenges of applying high-throughput sequencing in such a setting. The group enrolled patients with advanced or refractory cancer and performed whole-genome and RNA/transcriptome sequencing of the tumor, as well as targeted whole-exome sequencing of the tumor and normal DNA. They identified potentially informative mutations, including structural

rearrangements, CNVs, point mutations and gene-expression alterations in a clinically relevant time frame of 4 weeks. A multi-disciplinary Sequencing Tumor Board was then commissioned to provide clinical interpretation of the sequencing results obtained, and two of the patients were subsequently enrolled into clinical study protocols based on their cancer-genomic profile within 24 days of biopsy [66]. This is an early glimpse of the personalized medicine approach the authors will be able to adopt for cancer patients in the future. It is likely, however, that for the foreseeable future a combination of traditional techniques, including histology, immunohistochemistry, MC and FISH, will continue to be applied with increasing contributions from high-throughput technologies (FIGURE 2).

Biomarkers

The potential to develop personalized genomic biomarkers using NGS is particularly promising. These are patient-specific mutations that can be analyzed by NGS, PCR, quantitative PCR or digital PCR, and can be used in monitoring patient response to therapy. The personalized analysis of the rearranged ends method uses patient-specific rearrangements from tumor samples [67]. PCR assays can then be developed to detect rearrangements in normal blood, differentiating circulating tumor DNA from normal-genome equivalents. Researchers have developed these biomarkers in cancers including colorectal and breast, enabling mutant DNA molecules to be detected when present at a level of 0.001%. Limitations include the fact that rearrangements may be lost as a cancer evolves, and the clinical significance of such low levels of tumor DNA may be difficult to define. However,

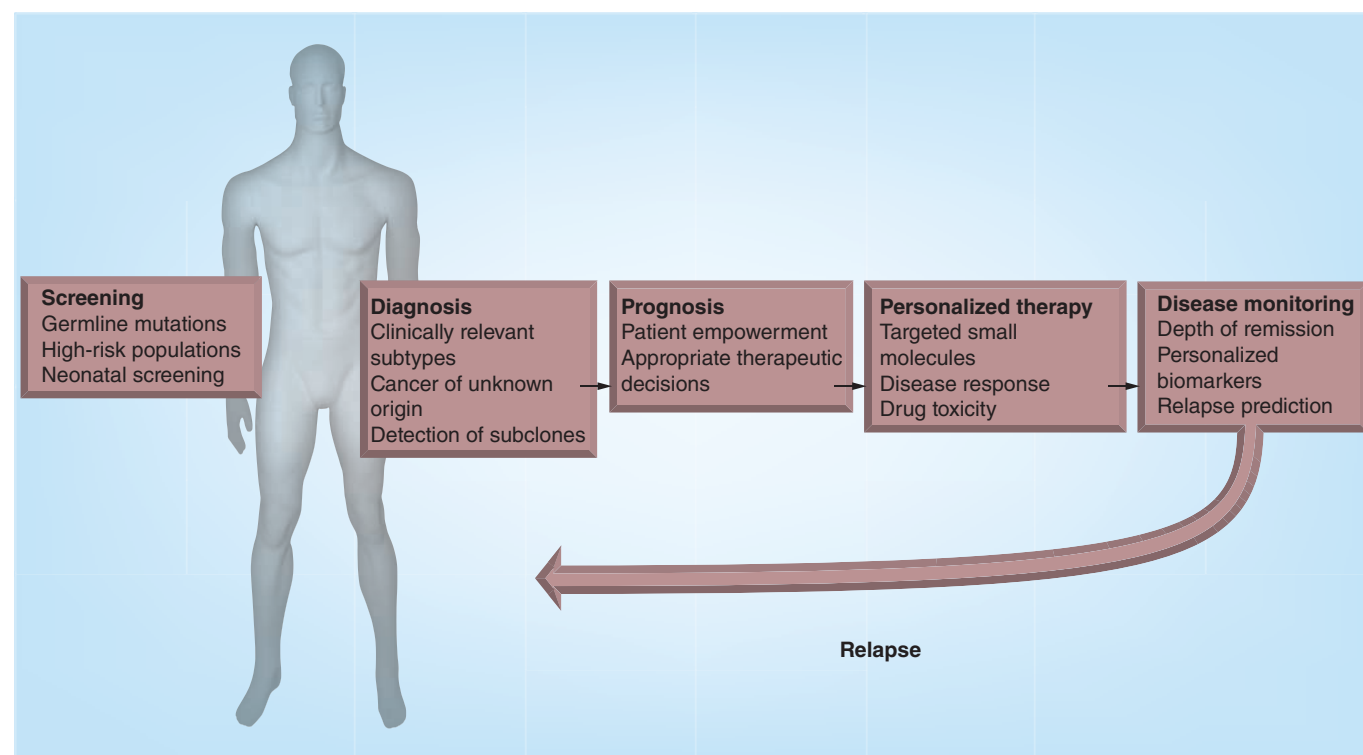


Figure 2. Role of high-throughput technologies in a patient's cancer journey.

a number of applications include accurately identifying surgical margins free of tumor, and assessing response to anticancer treatments. This is an area of clinical research where we are likely to see significant developments over the next 3 years.

Expert commentary

aCGH has been adopted by many clinical genetics laboratories as a first-line test for congenital abnormalities due to its automatability and potential to transform cancer genetics due to its detection-resolution for CNVs, deletions, amplifications, duplications and aneuploidies [68–70]. A range of aCGH platforms are commercially available and the individual clinical laboratory must match its requirements. Factors to be considered when choosing a platform include its resolution and sensitivity in addition to its cost and logistical aspects such as turnaround time and hands-on time. TABLE 2 summarizes some of these data for a range of available platforms. SNP-CGH is additionally useful for determining copy number-neutral rearrangements through LOH analysis, which can be used to infer tumor ploidy and stromal contamination using allele-specific copy-number analysis of tumors [71]. These factors mean that SNP arrays are a promising diagnostic, albeit currently more expensive, modality for clinical CGH applications in cancer. However, it must be appreciated that SNPs are not homogeneously distributed through the genome and the depth of coverage is thus not uniform. In addition, DNA derived from FFPE material is often unsuitable for this approach without significant technical and informatics adjustments. Array-based or SNP-CGH are also unable to detect balanced chromosomal translocations or inversions (for which FISH remains an important technique) and single basepair mutations and small insertions/deletions may also be missed [17].

Recent guidelines for the design and performance of clinical copy-number arrays have been released by the American College of Medical Genetics [72]. These state that arrays should use probe spacing that is compatible with detecting gains and losses of greater than 400 kb and SNP probes should be included where possible to provide supportive information for LOH applications. Manufacturers should also provide software to detect 99% of CNVs greater than 400 kb with a false-positive rate of less than 1%. The UK National Genetics Reference Laboratories recently compared five microarray platforms for identification of copy number aberrations [103]. They ran the same 12 cytogenetically abnormal samples and observed a core set of 15 abnormalities detected across the three arrays with highest probe density. This underscores the ACMG's findings that increased probe density is of fundamental importance. The rapid development of both aCGH and SNP-CGH technologies has meant that guidelines and comparisons are often made on a set of defined products. Comparisons have been made using clinical genetics samples and FFPE material [73,74]. Studies must be carried out very carefully if meaningful comparisons are to be made [75].

With the advent of highly accurate NGS assays with short run times and rapidly falling costs that are able to detect the full complement of genetic alterations found in cancer, there is no doubt that the longer term future of cancer genomics lies with

high-throughput sequencing technologies. A range of competing platforms are available, with Illumina technology currently providing the most popular due to its simplicity of sample preparation and ease of method development. Illumina combines short pieces of DNA with Y-shaped adapter oligos to create PCR amplifiable fragment libraries. Almost any DNA source can be used as a template including cDNA, meaning the number of available methods is very large [104]. In the last 5 years, the Illumina technology has improved from 1 Gb to over 1 Tb of data per run, and more recent technologies such as Ion Torrent have huge potential and may develop along a similar trajectory. The Illumina SBS chemistry is still improving, with almost 'Sanger length' reads of 700 bp being demonstrated in early 2012. When choosing a platform for sequencing cancer genomes, it is important to be aware of the depth of sequencing and physical coverage it offers, and to appreciate that NGS techniques can be affected by sequencing bias; for example, with poor coverage in areas of high GC content. To detect alterations reliably in a human genome sample of 3 billion bases requires at least 30-fold coverage on average (i.e., generation of 90 billion bases of sequence data per sample). For cancer samples, however, the depth of coverage needs to be increased to allow for increased ploidy, detection of cancer subclones and to account for the varying degrees of contamination with normal cells. This naturally carries a more substantial data load.

An alternative approach to simplify the use of NGS in a clinical context is to sequence only the exome or other specific regions of interest (ROI). An exome, or specific ROI, can be captured from a whole-genome sequencing library by hybridization to biotinylated 'bait' oligonucleotides. Methods were originally developed for solid-surface array-based capture but these have almost universally been replaced by in-solution solution-based capture [67,76–78]. Elution of only those library molecules that are bound to capture oligonucleotides creates a library with reduced complexity compared with the original whole genome library preparation. Several groups have compared the two different methods, and Mamanova *et al.* also compared these against PCR and molecular inversion probes [79]. New methods are being rapidly developed for targeted resequencing assays. For most clinical scenarios, the analysis of specific loci will prove faster, more cost-effective and require smaller amounts of nucleic acids for analysis and be significantly easier to interpret than whole-genome or exome analysis. There are increasing numbers of clinical molecular genetics laboratories adopting NGS for routine testing that are using PCR amplification as the method of choice for ROI selection [80]. PCR is sensitive and specific, allowing detection of rare alleles in complex and heterogeneous samples, even in circulating nucleic acid extracted from plasma [81]. The development of disease-specific NGS 'panels' may prove controversial, with differing opinions on what to include in any given panel. However, since many genes may play an as yet undiscovered role in a particular disease, panels should probably be as inclusive as technology allows while retaining the low-cost, short turn around time and ease of analysis amplicons bring. Smaller panels are likely to have an impact where larger numbers of samples are routinely available or in the context of screening assays. Very recently, Illumina released their TruSight™

cancer panel developed in conjunction with the Institute of Cancer Research and the Royal Marsden Hospital [105].

The year 2012 has seen the rapid uptake of lower yielding but faster 'personal' NGS sequencer instruments. The MiSeq from Illumina and PGM from Ion Torrent both allow sample preparation, sequencing and data analysis to be completed in under 24 h, opening the possibility for real-time clinical sequencing. Although these instruments are currently insufficient to provide whole-genome sequencing they are ideally suited to focused evaluation of specific cancer genes clinically useful for diagnosis, prognosis or therapy. New NGS technologies are under development, the most promising in the authors' opinion is the strand-sequencing approach being developed by Oxford Nanopore Technologies (Oxford, UK). They have coupled DNA polymerase to a nanopore, allowing label- and amplification-free sequencing. Although this is still in development, read lengths of up to 100,000 bp have been described and the possibility of a 20-min human genome sequence has been suggested although at an unknown cost. With the advent of such relatively inexpensive personal genome sequencers running highly accurate simple assays, there is no doubt that NGS technologies are poised ready to make their presence felt in clinical practice.

Five-year view

Over the last few years, the authors have witnessed a groundbreaking technological revolution, which has given us unprecedented access to the human germline and somatic genome. As with most branches of biology, cancer research has benefited greatly from such progress. As a result, researchers throughout the world have embarked on a systematic characterization of cancer genomes and are rapidly identifying the pathogenetic mutations underlying most forms of cancer. Over the next 5 years, the authors will have cataloged the complement of mutations for most forms of cancer, established their patterns of coexistence, and in many cases determined their clinical relevance. The International Cancer Genome Consortium is currently undertaking whole-genome analysis of 50 cancer subtypes that are of significant importance around the globe. Overall, the project aims to study over 25,000 cancer genomes at a genomic, epigenomic and transcriptomic level. The vast amount of information being generated by this international collaborative effort is made freely available through the Catalogue of Somatic Mutations in Cancer [82,106]. The next step in developing greater confidence with the use of NGS technology in clinical practice will arise through integration of sequencing data in sophisticated clinical trials that take into account the subtleties of phenotypic subdivisions based on individual patients' cancer-genomic features. This will, in time, lead to the development of a clinical grade database of clinically relevant cancer-associated mutations that influence significant therapeutic decisions, leading to improvements in patient care.

Sequence-based technologies are likely to dominate future cancer diagnostics, as they are capable of both incorporating and replacing current methodologies such as karyotyping and CGH, and more accurately predict phenotypic characteristics such as antigen expression, which are currently detected by flow cytometry

or immunohistochemistry. As sequencing technology becomes increasingly affordable, it will find application in all branches of pathology, including hematology, histopathology, microbiology and human genetics. Another advantage of sequence-based techniques is their unprecedented sensitivity, allowing detection of minimal residual disease. A recently developed protocol for single cancer-cell analysis, for example, may prove useful to study blood-borne or disseminated cancer cells in bone marrow and other organs that may persist after resection of the primary tumor. Molecular analysis of these cells may reveal unique information to tailor therapies and prevent seeding of metastases and subsequent relapse. Other groups are developing plasma-based tumor-DNA analysis methods. These could be used not only for making an initial diagnosis, but also as a form of minimal disease monitoring and personalized tumor markers, which may eventually be more robust and sensitive markers of disease progression than radiological detection of relapse. Such examples of intermediate scale of sequencing are most likely to have immediate impact on clinical genomics [67,81].

A number of challenges remain, however, before NGS technology, can play a central role in clinical cancer care. The authors growing understanding of tumor biology itself presents the greatest challenge. A cancer is an evolving unit of related malignant cells with significant tumor heterogeneity. Evaluation of biopsy material thus provides only a localized snapshot of the genetic features at a single point in a tumor's evolution. Even primary and metastatic tumors from the same patient may exhibit marked differences [83]. This has significant implications for both diagnostics and therapeutics [84]. The authors' appreciation of the role of epigenetic dysregulation in tumorigenesis is also deepening, and these changes would not currently be detected by either aCGH or NGS technologies [85]. Although high-throughput assays to detect epigenetic changes are unlikely to be incorporated into routine clinical practice in the near future, a recent study looking at integrated DNA copy number and methylation profiling of lymphoid neoplasms using a single Illumina Infinium Methylation assay demonstrated results that were comparable with using an Affymetrix 250K SNP array [86].

Although the costs of sequencing and data storage and processing have reduced significantly in recent years, and the speed and breadth of genome sequencing is increasing, the widespread use of such technology in routine clinical practice outside first-world teaching institutes may remain prohibitively expensive for a few years yet [107]. Targeted or exome-only sequencing approaches may be adopted more readily in clinical practice and prove more cost effective in the short to medium term. A recent study that exome-sequenced genomic alterations against eight cancer cell lines found 95% concordance with an Affymetrix SNP array and detected 19 out of 21 of the mutations reported in the Sanger COSMIC database for these lines, highlighting the feasibility of such an approach [87]. With current systems, the period of time required to generate sequencing data and to analyze, validate and interpret the results to produce a clinically actionable report may be clinically inappropriate for symptomatic patients requiring urgent treatment. Real-time NGS techniques such as the Oxford Nanopore Technology – although insufficient to provide whole-genome

sequencing currently – may be suitable for more focused evaluation of specific cancer genes clinically useful for diagnosis, prognosis or therapy. With the advent of such highly accurate, simple assays with short run times that can run on relatively small inexpensive instruments, there is no doubt that the technology itself has outpaced the methods of analysis and interpretation that are certainly the most significant current challenges to widespread use of NGS in clinical practice [28]. The computational resources required for assembly, annotation and analysis followed by clinically relevant interpretation are becoming the main bottlenecks. Novel algorithms for assembly and analysis will be vital, however, to prevent the volume of sequencing data overwhelming available resources. The challenge of interpreting genome-wide data is a challenge that cannot be understated, with driver and passenger mutations difficult to distinguish in a disordered, aberrant cancer genome. With the advent of benchtop sequencers that could be used routinely in hospital laboratories, specialized sequencing centers may begin to serve principally as bioinformatics resources that lend computational and interpretive resources and expertise to clinicians. Cancer genomics is not the only area where we can expect to see NGS advances being deployed in. Other fields can make use of the technologies and redefine clinical practice. The recent use of NGS in enhancing the detection of MRSA transmission in a hospital special care baby unit is an exciting demonstration of what is possible in other healthcare settings [88].

These developments will have profound implications on the future direction of research into the molecular pathogenesis and therapy of different cancers. Additionally, they have direct implications for the clinical management of individual patients and for the targeting of existing therapies. The format of the interaction between clinicians and outputs of diagnostic genomic data is hard to predict, but it is likely to involve a significant informatics component. What is clear is that national and international databases will be established for each cancer type, linking somatic mutations to disease phenotypes, drug responses, patient outcomes and probably a number of other variables. The genomic information derived from individual patients will be compared with one or

more curated databases, and this will generate not only diagnostic data but also prognostic information; advice on the most appropriate therapy; and, potentially, information regarding tumor etiology, drug toxicity and other clinically pertinent parameters. The extent of sequencing (e.g., targeted set of genes vs exome vs genome) will determine the breadth of the advice that can be provided. A sustained collaboration between research and clinical laboratories, and frontline and healthcare practitioners, will be vital to allow this to occur [89]. It will be essential to adopt a multidisciplinary approach toward cancer-genome analyses with oncologists, pathologists, geneticists, biostatisticians, bioethicists and policy-makers working together to meet the challenges and opportunities afforded by these new technologies and to allow rapid translation of the new biomarkers and therapeutic targets into routine clinical practice.

Finally, successful integration of these emerging technologies into clinical practice will necessitate greater public and patient understanding of their benefits and implications. These technologies pose a number of ethical issues that must be taken into consideration: from detection of genetic variants of unknown significance to counseling for clinically significant incidental findings, as well as data management and confidentiality issues. These factors will thus have specific implications for the traditional doctor–patient relationship and the way clinical consultations are conducted, requiring appropriate training of medical and nursing staff to allow the successful integration of these technologies into routine clinical practice, and leading to an acceleration of the momentum toward personalized cancer medicine and tangible benefits for patients [90,91].

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Key issues

- A range of clinically significant alterations occur in cancer genomes ranging from single point mutations to larger areas of copy-number variation or loss of heterozygosity. Reliable detection of such mutations within individual cancer genomes is necessary if we are to offer truly personalized cancer care to our patients.
- Current diagnostic methods including histopathology, immunohistochemistry, flow cytometry and FISH characterize tumors in a limited way. High-throughput technologies such as array-comparative genomic hybridization and next-generation sequencing provide unprecedented levels of information regarding individual tumors and can be exploited to characterize the tumor more fully.
- These technologies cannot only aid initial diagnosis, but also provide clinically valuable information about chemosensitivity and resistance, informing therapeutic decisions. In addition, they may provide information about prognosis and disease monitoring with personalized biomarkers.
- Developments in a range of competing next-generation sequencing platforms has lead to a dramatic reduction in cost, and increase in breadth and accuracy of sequencing. Although many technologies can be integrated into clinical services, it is likely that the future of cancer genomics will be underpinned by high-throughput sequencing technologies. Alternative technologies such as genome expression profiling and array-comparative genomic hybridization may, however, play an important role until the time when sequencing technology is readily available and interpretable.
- A range of biological, technical and socioethical challenges need to be addressed as sequencing technologies become fully integrated into clinical practice.

References

Papers of special note have been highlighted as:

• of interest

•• of considerable interest

- 1 Hochhaus A, Druker B, Sawyers C *et al.* Favorable long-term follow-up results over 6 years for response, survival, and safety with imatinib mesylate therapy in chronic-phase chronic myeloid leukemia after failure of interferon- α treatment. *Blood* 111(3), 1039–1043 (2008).
- 2 Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11(10), 685–696 (2010).
- 3 Ley TJ, Mardis ER, Ding L *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456(7218), 66–72 (2008).
- 4 Mardis ER, Ding L, Dooling DJ *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* 361(11), 1058–1066 (2009).
- 5 Pleasance ED, Cheetham RK, Stephens PJ *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463(7278), 191–196 (2010).
- 6 Pao W, Miller V, Zakowski M *et al.* EGF receptor gene mutations are common in lung cancers from ‘never smokers’ and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* 101(36), 13306–13311 (2004).
- 7 Tomlins SA, Rhodes DR, Perner S *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310(5748), 644–648 (2005).
- 8 Soda M, Choi YL, Enomoto M *et al.* Identification of the transforming *EML4-ALK* fusion gene in non-small-cell lung cancer. *Nature* 448(7153), 561–566 (2007).
- 9 Lapunzina P, Monk D. The consequences of uniparental disomy and copy number neutral loss-of-heterozygosity during human development and cancer. *Biol. Cell* 103(7), 303–317 (2011).
- 10 Wong DW, Leung EL, So KK *et al.*; University of Hong Kong Lung Cancer Study Group. The *EML4-ALK* fusion gene is involved in various histologic types of lung cancers from nonsmokers with wild-type *EGFR* and *KRAS*. *Cancer* 115(8), 1723–1733 (2009).
- 11 Kallioniemi A, Kallioniemi OP, Sudar D *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258(5083), 818–821 (1992).
- 12 Chen X, Knauf JA, Gonsky R *et al.* From amplification to gene in thyroid cancer: a high-resolution mapped bacterial-artificial-chromosome resource for cancer chromosome aberrations guides gene discovery after comparative genome hybridization. *Am. J. Hum. Genet.* 63(2), 625–637 (1998).
- 13 Pinkel D, Seagraves R, Sudar D *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20(2), 207–211 (1998).
- 14 Wicker N, Carles A, Mills IG *et al.* A new look towards BAC-based array CGH through a comprehensive comparison with oligo-based array CGH. *BMC Genomics* 8, 84 (2007).
- 15 Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.* 57(6), 644–646 (2004).
- 16 Le Scouarnec S, Gribble SM. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. *Heredity (Edinb.)* 108(1), 75–85 (2012).
- 17 Maciejewski JP, Tiu RV, O’Keefe C. Application of array-based whole genome scanning technologies as a cytogenetic tool in haematological malignancies. *Br. J. Haematol.* 146(5), 479–488 (2009).
- 18 LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 37(13), 4181–4193 (2009).
- 19 Peiffer DA, Le JM, Steemers FJ *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* 16(9), 1136–1148 (2006).
- 20 Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4), 557–572 (2004).
- 21 Campbell PJ, Stephens PJ, Pleasance ED *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* 40(6), 722–729 (2008).
- 22 Venter JC, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science* 291(5507), 1304–1351 (2001).
- 23 Metzker ML. Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11(1), 31–46 (2010).
- This comprehensive review of next-generation sequencing (NGS) technologies covers all the major points such as DNA library construction, template preparation by clustering or emulsion PCR, and the different sequencing chemistries, as well as genome enrichment and applications of NGS. As it is an earlier review, it is missing the Ion Torrent™ technology; however, it is still the most comprehensive.
- 24 Margulies M, Egholm M, Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057), 376–380 (2005).
- First 454 paper and arguably the first NGS paper.
- 25 Wheeler DA, Srinivasan M, Egholm M *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452(7189), 872–876 (2008).
- 26 Pleasance ED, Stephens PJ, O’Meara S *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463(7278), 184–190 (2010).
- 27 Bentley DR, Balasubramanian S, Swerdlow HP *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218), 53–59 (2008).
- First paper describing whole-genome Illumina sequencing.
- 28 Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol.* 12(8), 125 (2011).
- Determining the cost of human genome sequencing is a ‘how long is a piece of string’ question: the answer is almost always ‘it depends’. The authors of this paper have suggested four steps in genome data generation that the field should focus on when trying to determine or compare costs. We have used their formula to determine costs of sequencing in the corresponding author’s laboratory at the time of publication.
- 29 Loman NJ, Misra RV, Dallman TJ *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30(5), 434–439 (2012).
- 30 Bell CJ, Dinwiddie DL, Miller NA *et al.* Carrier testing for severe childhood recessive diseases by next-generation

- sequencing. *Sci. Transl. Med.* 3(65), 65ra4 (2011).
- 31 King MC, Marks JH, Mandell JB; New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in *BRCA1* and *BRCA2*. *Science* 302(5645), 643–646 (2003).
 - 32 Alsop K, Fereday S, Meldrum C *et al.* *BRCA* mutation frequency and patterns of treatment response in *BRCA* mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. *J. Clin. Oncol.* 30(21), 2654–2663 (2012).
 - 33 Walsh T, Lee MK, Casadei S *et al.* Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc. Natl Acad. Sci. USA* 107(28), 12629–12633 (2010).
 - 34 Tagawa H, Suguro M, Tsuzuki S *et al.* Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. *Blood* 106(5), 1770–1777 (2005).
 - 35 Seto M. Genomic profiles in B cell lymphoma. *Int. J. Hematol.* 92(2), 238–245 (2010).
 - 36 Takeuchi I, Tagawa H, Tsujikawa A *et al.* The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of B-cell lymphomas and genetic regions involved in lymphomagenesis. *Haematologica* 94(1), 61–69 (2009).
 - 37 Sehn LH. Early detection of patients with poor risk diffuse large B-cell lymphoma. *Leuk. Lymphoma* 50(11), 1744–1747 (2009).
 - 38 Robledo C, García JL, Caballero D *et al.*; Spanish Lymphoma/Autologous Bone Marrow Transplant Study Group (GEL-TAMO). Array comparative genomic hybridization identifies genetic regions associated with outcome in aggressive diffuse large B-cell lymphomas. *Cancer* 115(16), 3728–3737 (2009).
 - 39 Kreisel F, Kulkarni S, Kerns RT *et al.* High resolution array comparative genomic hybridization identifies copy number alterations in diffuse large B-cell lymphoma that predict response to immuno-chemotherapy. *Cancer Genet.* 204(3), 129–137 (2011).
 - 40 Barrans S, Crouch S, Smith A *et al.* Rearrangement of *MYC* is associated with poor prognosis in patients with diffuse large B-cell lymphoma treated in the era of rituximab. *J. Clin. Oncol.* 28(20), 3360–3365 (2010).
 - 41 Hagenkord JM, Monzon FA, Kash SF, Lilleberg S, Xie Q, Kant JA. Array-based karyotyping for prognostic assessment in chronic lymphocytic leukemia: performance comparison of Affymetrix 10K2.0, 250K Nsp, and SNP6.0 arrays. *J. Mol. Diagn.* 12(2), 184–196 (2010).
 - 42 Tybäckinoja A, Elonen E, Piippo K, Porkka K, Knuutila S. Oligonucleotide array-CGH reveals cryptic gene copy number alterations in karyotypically normal acute myeloid leukemia. *Leukemia* 21(3), 571–574 (2007).
 - 43 O'Keefe CL, Tiu R, Gondek LP *et al.* High-resolution genomic arrays facilitate detection of novel cryptic chromosomal lesions in myelodysplastic syndromes. *Exp. Hematol.* 35(2), 240–251 (2007).
 - 44 Starczynowski DT, Vercauteren S, Telenius A *et al.* High-resolution whole genome tiling path array CGH analysis of CD34⁺ cells from patients with low-risk myelodysplastic syndromes reveals cryptic copy number alterations and predicts overall and leukemia-free survival. *Blood* 112(8), 3412–3424 (2008).
 - 45 Tiu RV, Gondek LP, O'Keefe CL *et al.* New lesions detected by single nucleotide polymorphism array-based chromosomal analysis have important clinical impact in acute myeloid leukemia. *J. Clin. Oncol.* 27(31), 5219–5226 (2009).
 - 46 Gondek LP, Tiu R, O'Keefe CL, Sekeres MA, Theil KS, Maciejewski JP. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. *Blood* 111(3), 1534–1542 (2008).
 - 47 Tiu RV, Gondek LP, O'Keefe CL *et al.* Prognostic impact of SNP array karyotyping in myelodysplastic syndromes and related myeloid malignancies. *Blood* 117(17), 4552–4560 (2011).
 - 48 Makishima H, Rataul M, Gondek LP *et al.* FISH and SNP-A karyotyping in myelodysplastic syndromes: improving cytogenetic detection of del(5q), monosomy 7, del(7q), trisomy 8 and del(20q). *Leuk. Res.* 34(4), 447–453 (2010).
 - 49 Welch JS, Westervelt P, Ding L *et al.* Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* 305(15), 1577–1584 (2011).
 - 50 Massard C, Lorient Y, Fizazi K. Carcinomas of an unknown primary origin – diagnosis and treatment. *Nat. Rev. Clin. Oncol.* 8(12), 701–710 (2011).
 - 51 Evers B, Drost R, Schut E *et al.* Selective inhibition of *BRCA2*-deficient mammary tumor cell growth by AZD2281 and cisplatin. *Clin. Cancer Res.* 14(12), 3916–3925 (2008).
 - 52 Shaw AT, Yeap BY, Solomon BJ *et al.* Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring *ALK* gene rearrangement: a retrospective analysis. *Lancet Oncol.* 12(11), 1004–1012 (2011).
 - 53 Lièvre A, Bachet JB, Le Corre D *et al.* *KRAS* mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res.* 66(8), 3992–3995 (2006).
 - 54 Molinari F, Frattini M. *KRAS* mutational test for metastatic colorectal cancer patients: not just a technical problem. *Expert Rev. Mol. Diagn.* 12(2), 123–126 (2012).
 - 55 Pirker R, Herth FJ, Kerr KM *et al.*; European EGFR Workshop Group. Consensus for *EGFR* mutation testing in non-small cell lung cancer: results from a European workshop. *J. Thorac. Oncol.* 5(10), 1706–1713 (2010).
 - 56 Lipson D, Capelletti M, Yelensky R *et al.* Identification of new *ALK* and *RET* gene fusions from colorectal and lung cancer biopsies. *Nat. Med.* 18(3), 382–384 (2012).
 - 57 Beroukhi R, Mermel CH, Porter D *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* 463(7283), 899–905 (2010).
 - 58 Sequist LV, Heist RS, Shaw AT *et al.* Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. *Ann. Oncol.* 22(12), 2616–2624 (2011).
 - 59 Davies H, Bignell GR, Cox C *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* 417(6892), 949–954 (2002).
 - 60 Sosman JA, Kim KB, Schuchter L *et al.* Survival in *BRAF* V600-mutant advanced melanoma treated with vemurafenib. *N. Engl. J. Med.* 366(8), 707–714 (2012).
 - 61 Fedorenko IV, Paraiso KH, Smalley KS. Acquired and intrinsic *BRAF* inhibitor resistance in *BRAF* V600E mutant melanoma. *Biochem. Pharmacol.* 82(3), 201–209 (2011).
 - 62 Nazarian R, Shi H, Wang Q *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468(7326), 973–977 (2010).
 - 63 Wagle N, Emery C, Berger MF *et al.* Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic

- profiling. *J. Clin. Oncol.* 29(22), 3085–3096 (2011).
- 64 West C, Rosenstein BS, Alsner J *et al.*; EQUAL-ESTRO. Establishment of a Radiogenomics Consortium. *Int. J. Radiat. Oncol. Biol. Phys.* 76(5), 1295–1296 (2010).
 - 65 Barnett GC, Coles CE, Elliott RM *et al.* Independent validation of genes and polymorphisms reported to be associated with radiation toxicity: a prospective analysis study. *Lancet Oncol.* 13(1), 65–77 (2012).
 - 66 Roychowdhury S, Iyer MK, Robinson DR *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* 3(111), 111ra121 (2011).
 - 67 Leary RJ, Kinde I, Diehl F *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* 2(20), 20ra14 (2010).
 - 68 Xiang B, Zhu H, Shen Y *et al.* Genome-wide oligonucleotide array comparative genomic hybridization for etiological diagnosis of mental retardation: a multicenter experience of 1499 clinical cases. *J. Mol. Diagn.* 12(2), 204–212 (2010).
 - 69 Lichtenbelt KD, Knoers NV, Schuring-Blom GH. From karyotyping to array-CGH in prenatal diagnosis. *Cytogenet. Genome Res.* 135(3–4), 241–250 (2011).
 - 70 Shaffer LG, Bejjani BA. Medical applications of array CGH and the transformation of clinical cytogenetics. *Cytogenet. Genome Res.* 115(3–4), 303–309 (2006).
 - 71 Van Loo P, Nordgard SH, Lingjærde OC *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* 107(39), 16910–16915 (2010).
 - 72 Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST; Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* 13(7), 680–685 (2011).
 - 73 Krijgsman O, Israeli D, Haan JC *et al.* CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material. *Genes. Chromosomes Cancer* 51(4), 344–352 (2012).
 - 74 Pinto D, Darvishi K, Shi X *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* 29(6), 512–520 (2011).
 - 75 Curtis C, Lynch AG, Dunning MJ *et al.* The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* 10, 588 (2009).
 - 76 Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. *Am. J. Clin. Pathol.* 136(4), 527–539 (2011).
 - 77 Bainbridge MN, Wang M, Burgess DL *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol.* 11(6), R62 (2010).
 - 78 Albert TJ, Molla MN, Muzny DM *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4(11), 903–905 (2007).
 - 79 Mamanova L, Coffey AJ, Scott CE *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7(2), 111–118 (2010).
- The choice of method for targeting specific regions of interest for NGS varies from laboratory to laboratory. This paper compares the major methods currently in use and was the first to suggest precapture pooling, which is now available in most commercially available kits. The authors are firmly in favor of hybrid selection over array-based, molecular inversion probe (MIP) or PCR methods. PCR is given a particularly poor rating and unfortunately the authors do not address any of the next-generation PCR methods that are becoming increasingly widely adopted.
- 80 Morgan JE, Carr IM, Sheridan E *et al.* Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum. Mutat.* 31(4), 484–491 (2010).
 - 81 Forshew T, Murtaza M, Parkinson C *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* 4(136), 136ra68 (2012).
- First demonstration of how tumor DNA circulating in a patient's blood can be used to monitor disease and opens the possibility of earlier detection of relapse.
- 82 Forbes SA, Bhamra G, Bamford S *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* 10, Unit 10.11 (2008).
- Databases such as COSMIC will become increasingly important in the development of clinical tests. We are likely to require more and more data in the public domain to make sense of what could be termed 'private' mutations in smaller studies. Annotation of databases with clinically relevant information will be a focus of much debate and action over the next few years.
- 83 Poplawski AB, Jankowski M, Erickson SW *et al.* Frequent genetic differences between matched primary and metastatic breast cancer provide an approach to identification of biomarkers for disease progression. *Eur. J. Hum. Genet.* 18(5), 560–568 (2010).
 - 84 Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* 12(5), 323–334 (2012).
 - 85 Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 31(1), 27–36 (2010).
 - 86 Kwee I, Rinaldi A, Rancoita P *et al.* Integrated DNA copy number and methylation profiling of lymphoid neoplasms using a single array. *Br. J. Haematol.* 156(3), 354–357 (2012).
 - 87 Chang H, Jackson DG, Kayne PS, Ross-Macdonald PB, Ryseck RP, Siemers NO. Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS ONE* 6(6), e21097 (2011).
 - 90 Harris SR, Cartwright EJ, Török ME *et al.* Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* 13(2), 130–136 (2013).
- Demonstrates the utility of public health surveillance using NGS technologies. The methods described are likely to appear in molecular genetics laboratories for specific cancer gene resequencing, with turnaround times of significantly less than the 2 weeks that patients can expect to wait for referral to an oncologist.
- 91 Desai AN, Jere A. Next-generation sequencing: ready for the clinics? *Clin. Genet.* 81(6), 503–510 (2012).
 - 92 Hastings R, de Wert G, Fowler B *et al.* The changing landscape of genetic testing and its impact on clinical and laboratory services and research in Europe. *Eur. J. Hum. Genet.* 20(9), 911–916 (2012).
 - 91 Ong FS, Grody WW, Deignan JL. Privacy and data management in the era of massively parallel next-generation sequencing. *Expert Rev. Mol. Diagn.* 11(5), 457–459 (2011).

Websites

- 101 NICE NHS. Familial breast cancer. www.nice.org.uk/nicemedia/pdf/CG41NICEguidance.pdf

- 102 BRACAnalysis® Technical Specifications. Myriad Genetic Laboratories, Updated April 2012. www.myriad.com/lib/technical-specifications/BRACAnalysis-Technical-Specifications.pdf
- 103 Huang SCJ. Comparison for cytogenetics array platforms hardware and software for use in identifying copy number aberrations in constitutional disorders (2010). www.ngsl.org.uk/Wessex/downloads_reports.htm
- 104 CoreGenomics. <http://core-genomics.blogspot.co.uk/2011/09/next-generation-sequencing-acronyms.html>
- 105 Illumina. TrueSight™ Cancer. www.illumina.com/Documents/products/datasheets/datasheet_TrueSight_Cancer.pdf
- 106 Wellcome Trust. Sanger Institute. www.sanger.ac.uk/genetics/CGP/cosmic
- 107 NIH. National Human Genome Research Institute. DNA Sequencing Costs. www.genome.gov/sequencingcosts

Somatic mutations in *ATP1A1* and *CACNA1D* underlie a common subtype of adrenal hypertension

Elena A B Azizan^{1,12}, Hanne Poulsen^{2,12}, Petronel Tuluc^{3,12}, Junhua Zhou^{1,12}, Michael V Clausen², Andreas Lieb³, Carmela Maniero¹, Sumedha Garg⁴, Elena G Bochukova⁴, Wanfeng Zhao⁵, Lalarukh Haris Shaikh¹, Cheryl A Brighton¹, Ada E D Teo¹, Anthony P Davenport¹, Tanja Dekkers⁶, Bas Tops⁷, Benno Küsters⁷, Jiri Ceral⁸, Giles S H Yeo⁴, Sudeshna Guha Neogi⁹, Ian McFarlane⁹, Nitzan Rosenfeld¹⁰, Francesco Marass¹⁰, James Hadfield¹⁰, Wojciech Margas¹¹, Kanchan Chaggar¹¹, Miroslav Solar⁸, Jaap Deinum⁶, Annette C Dolphin¹¹, I Sadaf Farooqi^{4,12}, Joerg Striessnig^{3,12}, Poul Nissen^{2,12} & Morris J Brown^{1,12}

At least 5% of individuals with hypertension have adrenal aldosterone-producing adenomas (APAs). Gain-of-function mutations in *KCNJ5* and apparent loss-of-function mutations in *ATP1A1* and *ATP2A3* were reported to occur in APAs^{1,2}. We find that *KCNJ5* mutations are common in APAs resembling cortisol-secreting cells of the adrenal zona fasciculata but are absent in a subset of APAs resembling the aldosterone-secreting cells of the adrenal zona glomerulosa³. We performed exome sequencing of ten zona glomerulosa-like APAs and identified nine with somatic mutations in either *ATP1A1*, encoding the Na⁺/K⁺ ATPase α 1 subunit, or *CACNA1D*, encoding Ca_v1.3. The *ATP1A1* mutations all caused inward leak currents under physiological conditions, and the *CACNA1D* mutations induced a shift of voltage-dependent gating to more negative voltages, suppressed inactivation or increased currents. Many APAs with these mutations were <1 cm in diameter and had been overlooked on conventional adrenal imaging. Recognition of the distinct genotype and phenotype for this subset of APAs could facilitate diagnosis.

APAs are the most common curable cause of hypertension^{4,5} and are often due to specific somatic mutations^{1,2}. Gain-of-function mutations in the potassium channel *KCNJ5* were found in approximately 40% of APAs^{1,3,6,7}, and mutations in *ATP1A1* and *ATP2A3*, two P-type ATPases regulating Na⁺, K⁺ and Ca²⁺ transport, were recently discovered in a further 7% of APAs². Here we report mutations in two genes regulating Na⁺, K⁺ and Ca²⁺ transport (*ATP1A1* and *CACNA1D*) and highlight the existence of distinct APA subtypes with different

mutation profiles. Functional studies of these mutations provide explanations for their dominant effects.

We looked for somatic mutations in APAs with a zona glomerulosa-like phenotype. The zona glomerulosa is the principal site of aldosterone secretion and cell turnover in the adrenal gland, but, paradoxically, classical 'Conn's tumors' tend to resemble cells of the cortisol-secreting

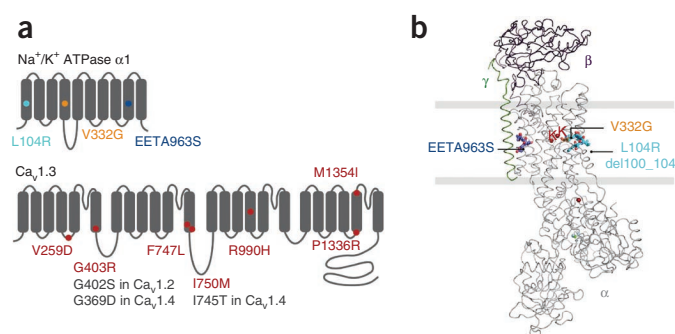


Figure 1 Somatic mutations in *ATP1A1* and *CACNA1D* in APAs. (a) Schematic of Na⁺/K⁺ ATPase subunit α 1 and Ca_v1.3. Colored circles indicate the positions where somatic alterations or deletions have been described in APAs. (b) An overview of the E2.Pi Na⁺/K⁺ ATPase showing the tripartite complex of α (gray), β (purple) and γ (green) subunits with the extracellular space on top and the membrane represented by two horizontal gray lines. The two occluded K⁺ molecules (red) and the substitutions and deletions identified in APAs (colored as in a) are indicated. The image was generated with PyMOL using Protein Data Bank (PDB) 2ZXE.

¹Clinical Pharmacology Unit, Centre for Clinical Investigation, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK. ²Centre for Membrane Pumps in Cells and Disease—PUMPKin, Danish National Research Foundation, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. ³Pharmacology and Toxicology, Institute of Pharmacy, Center for Molecular Biosciences, University of Innsbruck, Innsbruck, Austria. ⁴University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. ⁵Human Research Tissue Bank, Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Addenbrooke's Hospital, Cambridge, UK. ⁶Department of Internal Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. ⁷Department of Pathology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. ⁸1st Department of Internal Medicine—Cardioangiopathy, Charles University Faculty of Medicine in Hradec Kralove and University Hospital Hradec Kralove, Hradec Kralove, Czech Republic. ⁹Cambridge National Institute for Health Research (NIHR) Biomedical Research Centre (BRC), Department of Clinical Biochemistry, Addenbrooke's Hospital, Cambridge, UK. ¹⁰Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK. ¹¹Department of Neuroscience, Physiology and Pharmacology, University College London, London, UK. ¹²These authors contributed equally to this work. Correspondence should be addressed to M.J.B. (m.j.brown@cai.cam.ac.uk) or H.P. (hp@mb.au.dk).

Received 4 March; accepted 3 July; published online 4 August 2013; doi:10.1038/ng.2716

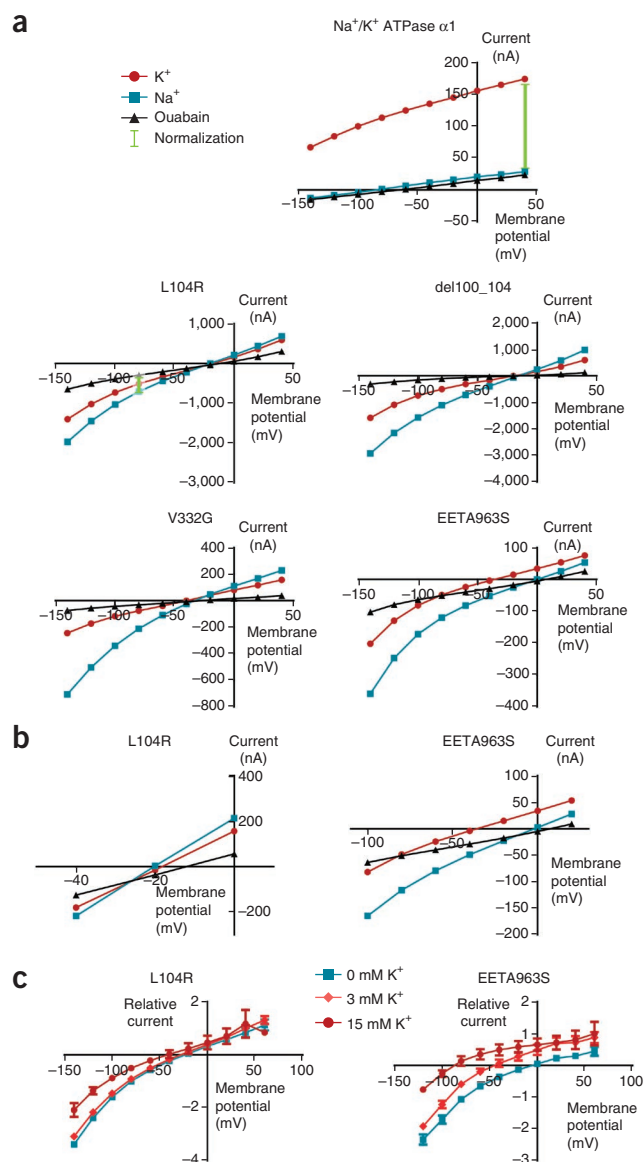
Figure 3 Gain-of-function alterations in the Na⁺/K⁺ ATPase cause inward current under physiological conditions. **(a)** Examples of the raw currents measured in individual oocytes expressing the human Na⁺/K⁺ ATPase with $\alpha 1$, either wild type or with one of the four alterations identified. Representative of 6–22 measurements. In wild-type pumps, extracellular K⁺ activates forward pumping, with maximal current generated at positive membrane potentials, and that value is usually used for normalization. In the absence of extracellular K⁺, the difference between currents in Na⁺ with and without ouabain is small. In contrast, the four mutants identified in APAs have distinct inward, ouabain-sensitive currents in Na⁺. To pool measurements from different oocytes, normalization of the mutant data therefore has to be carried out fundamentally differently. We used the inward leak in Na⁺, namely subtracting out the ouabain-sensitive current measured in K⁺-free buffer at –80 mV. **(b)** Enlarged view of the current curves showing that extracellular K⁺ activates an outward current with GluGluThrAla963Ser but not with Leu104Arg mutant channel. For the Leu104Arg mutant, extracellular K⁺ is a competitive, non-conducting inhibitor, suggesting that the leaking pumps are in a conformation open to the outside, where they are able to bind K⁺ (which competes with leaking) but unable to occlude K⁺ and therefore unable to pump. **(c)** Extracellular K⁺ inhibits the leak current in a dose-dependent manner. Leu104Arg mutant: error bars, s.e.m.; $n = 3$ –7 oocytes; average normalization value = 452 nA. GluGluThrAla963Ser mutant: error bars, s.e.m.; $n = 3$ –8 oocytes; average normalization value = 94 nA.

and **Supplementary Table 1b,c**). The tenth zona glomerulosa-like APA had a *CTNNB1* substitution mutation encoding p.Ser33Cys. None of the ten zona glomerulosa-like APAs had *KCNJ5* mutations, and none of the *ATP1A1* or *CACNA1D* somatic mutations were found in 100 healthy, normotensive individuals or in 8,000 publicly available exomes.

To confirm that APAs are associated with *ATP1A1* and *CACNA1D* somatic mutations, we screened our remaining Cambridge cohort and two independently ascertained cohorts, a Dutch and a Czech cohort (where visible nodules on adrenal imaging were not a prerequisite for inclusion). We genotyped all three cohorts for the five substitution mutations in *ATP1A1* and *CACNA1D* and identified deletions and additional substitutions through microfluidic sequencing of *ATP1A1* and *CACNA1D* exons. We found the *ATP1A1* mutation encoding p.Leu104Arg in 3 of 53 APAs, 4 of 39 adenomas and 1 of 91 nodular lesions (1/50 subjects) in our remaining Cambridge, Czech and Dutch cohorts, respectively. No further deletion mutations (of *ATP1A1*) were found in either the Dutch or Czech cohort. We also found three of the *CACNA1D* mutations in these cohorts, together with three previously undescribed mutations of conserved residues, one of which occurred twice (7/142 subjects) (**Supplementary Figs. 1 and 2** and **Supplementary Table 2**).

The phenotype of APAs with the new mutations was compared with that of *KCNJ5*-mutant APAs. Individuals with the newly discovered mutations were mainly older males; the APAs were smaller (some ≤ 0.5 cm in diameter) with more compact, zona glomerulosa-like cells. In the Cambridge cohort, clinical presentation, immunohistochemistry for the gene products, the transcriptome and the presence of spironolactone bodies differed between genotypes (**Fig. 2** and **Supplementary Figs. 3–6**). Among the genes upregulated in zona glomerulosa-like APAs with *ATP1A1* or *CACNA1D* mutations, several also showed higher expression in normal zona glomerulosa than in zona fasciculata³ (**Fig. 2** and **Supplementary Fig. 4**).

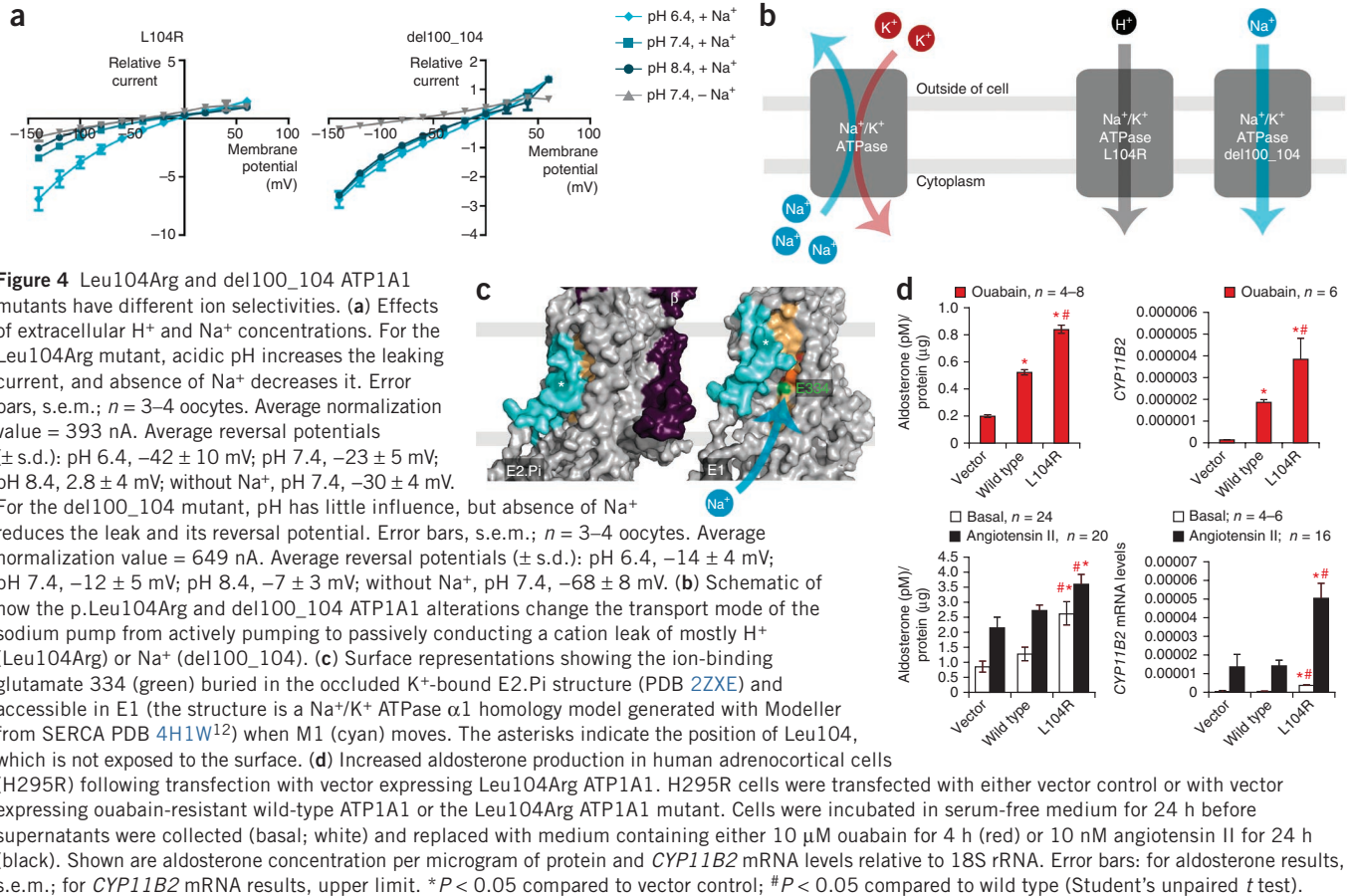
We tested the functional effects of the *ATP1A1* mutations by expression of mutant constructs in *Xenopus laevis* oocytes (**Fig. 3**). The Leu104Arg, del100_104 and Val332Gly² mutants showed no K⁺-stimulated pumping, but, at physiological membrane potentials, they all caused a marked ouabain-sensitive and voltage-dependent inward current that was partly inhibited by K⁺ (**Fig. 3a,c**). Although



the GluGluThrAla963Ser mutant did respond to K⁺ with forward pumping, it also caused a ouabain-sensitive inward current (**Fig. 3a,b**), and, under physiologically relevant conditions (–80 mV and 3 mM extracellular K⁺), the net contribution of the GluGluThrAla963Ser mutant was an inward current (**Fig. 3c**).

To determine which ions carry the currents, we varied the extracellular pH and exchanged Na⁺ for *N*-methyl-D-glucamine (NMDG) for the Leu104Arg and del100_104 mutants. For the Leu104Arg mutant, a change in pH of 1 changed the reversal potential by about 20 mV, whereas removal of extracellular Na⁺ had little effect on the reversal potential, suggesting that protons are the main carrier of the current. In contrast, the del100_104 mutant responded only a little to pH changes, whereas Na⁺ removal shifted the reversal potential by about 50 mV, suggesting that Na⁺ is the main carrier of current for the del100_104 mutant (**Fig. 4a,b**). Because both the Leu104Arg and del100_104 mutants are linked to zona glomerulosa-like APAs, it is likely that pathology is due to a common downstream effect of the Na⁺/K⁺ ATPase becoming permeable to either Na⁺ or protons.

Na⁺/K⁺ ATPases and the related Ca²⁺ ATPases have a highly conserved leucine in the middle of transmembrane helix 1 (M1) at the



position corresponding to Leu104 in Na^+/K^+ ATPase subunit $\alpha 1$. The crystal structure of the Na^+/K^+ ATPase shows that Leu104 is only 4 Å from a key ion-binding residue in M4, Glu334, at the so-called site II that is important for both Na^+ and K^+ binding (Fig. 1b)¹¹. For the closely related Ca^{2+} ATPase SERCA, a recent structure shows how Ca^{2+} entry (corresponding to Na^+ entry in the Na^+/K^+ ATPase) depends on a 12-Å sliding of M1 relative to M4 within the membrane, creating an opening toward site II (Fig. 4c)^{12,13}. The p.Leu104Arg and p.del100_104 alterations as well as the p.Val332Gly substitution² are all expected to affect the sliding of M1 and the structural context of Glu334 (Figs. 1b and 4c). In contrast, the p.GluGluThrAla963Ser substitution is on the other side of the transmembrane domain and includes Glu961, a glutamate of debated importance for the Na^+ -specific site III (Fig. 1b)^{14,15}, and the GluGluThrAla963Ser mutant responds differently to K^+ , so its leaking mechanism may differ from those identified for the other mutants. Unlike a previously described inwardly rectifying proton leak in wild-type sodium pumps¹⁵⁻¹⁷, the strong leaks created by the Na^+/K^+ ATPase $\alpha 1$ mutants identified here are present at physiological potentials and concentrations of Na^+ and K^+ (Fig. 4).

No major germline mutation of *ATP1A1* has been described, suggesting that loss of function would be either incompatible with life² or harmless. From our finding that all *ATP1A1* mutations identified in APAs cause inward leak currents, we infer that their heterozygous loss of pumping activity is not sufficiently deleterious to cause APAs. Notably, concentrations of ouabain that inhibit substantially more than half of the pump activity either inhibit or cause a transient increase in aldosterone secretion¹⁸⁻²¹. Increased aldosterone synthesis and secretion by human adrenocortical (H295R) cells transfected with

vector encoding ouabain-resistant Leu104Arg ATP1A1 (compared to vector encoding wild-type protein or empty vector) was unaffected by blockade of endogenous Na^+/K^+ ATPases (Fig. 4d). This finding seems to implicate gain of function rather than haploinsufficiency; however, both consequences of these mutations could be important, and further experiments with gene knockdown will be of interest.

Six of the seven newly discovered somatic mutations of *CACNA1D* encoding $Ca_v1.3$ affect conserved sites within functional domains known to form the channel activation gate (p.Gly403Arg, p.Ile750Met and p.Phe747Leu), the voltage sensor (p.Arg990His) and the cytoplasmic S4-S5 linker coupling the voltage-sensing domain to the pore (p.Pro1336Arg and p.Val259Asp) (Fig. 1b and Supplementary Fig. 2). Disease-related mutations at positions corresponding to p.Gly403Arg and p.Ile750Met in closely related $\alpha 1$ subunits (Fig. 1b) are known to affect channel function by shifting voltage dependence to more negative potentials and/or by slowing voltage-dependent inactivation²²⁻²⁴. Expression of the initial four $Ca_v1.3$ mutants in a human embryonic kidney cell line (tsA-201), together with accessory $\beta 3$ (unless otherwise stated) and $\alpha 2\delta 1$ subunits indeed resulted in gating changes suggestive of a gain-of-function phenotype. The Val259Asp and Ile750Met mutants shifted voltage-dependent activation and steady-state inactivation of $Ca_v1.3$ -mediated inward Ca^{2+} currents (I_{Ca}) by 15 mV to more negative voltages, independent of the concentration of charge carrier (15 or 2 mM extracellular Ca^{2+}) (Fig. 5a,b and Supplementary Table 3). The Val259Asp, Ile750Met and Pro1336Arg substitution mutants also slowed the biexponential I_{Ca} inactivation time course during prolonged (5-s) depolarizations (Fig. 5c and Supplementary Table 4). They also increased the ratio of tail current amplitude (I_{tail}) to integrated ON gating charge (Q_{ON} ,

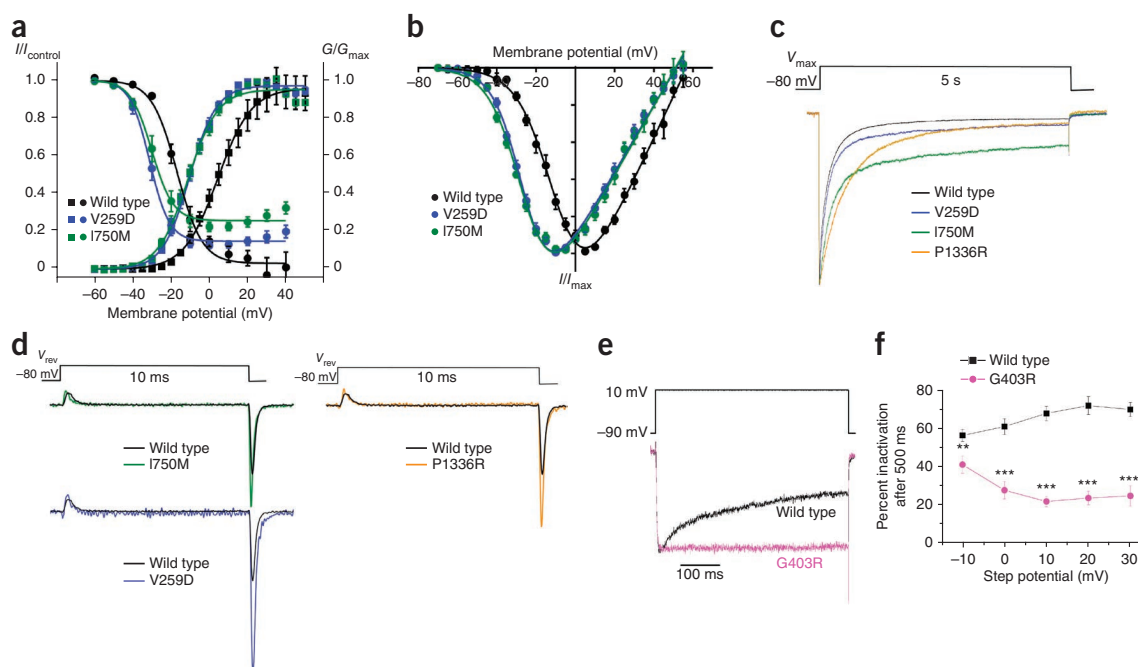


Figure 5 Functional consequences of *CACNA1D* mutations on $\text{Ca}_v1.3$ channel function. (a) Voltage dependence of activation (G/G_{max}) and steady-state inactivation (I/I_{control}) of wild-type $\text{Ca}_v1.3$ ($n = 11$) and $\text{Ca}_v1.3$ mutants Val259Asp ($n = 10$) and Ile750Met ($n = 11$) expressed in tsA-201 cells (expressed with $\beta 3$ and $\alpha 2\delta 1$, 15 mM Ca^{2+} as charge carrier). Activation data were calculated from current-voltage relationships (Online Methods). (b) Current-voltage relationship of mutants Val259Asp ($n = 9$) and Ile750Met ($n = 9$) (2 mM Ca^{2+} charge carrier). No changes in voltage-dependent gating were observed for mutant Pro1336Arg (activation, $n = 10$; inactivation, $n = 8$), and it was therefore omitted from the graphs. Fitted parameters for Pro1336Arg and other mutants are given in **Supplementary Table 3**. (c) I_{Ca} (15 mM Ca^{2+}) inactivation during 5-s depolarizations to V_{max} . Representative normalized currents are shown for wild-type channels and the indicated mutants. Double-exponential inactivation time-course statistics are given in **Supplementary Table 4** (wild type, $n = 9$; Val259Asp, $n = 9$; Ile750Met, $n = 9$; Pro1336Arg, $n = 10$). (d) Effects of $\text{Ca}_v1.3$ alterations (Val259Asp, $n = 10$; Ile750Met, $n = 8$; Pro1336Arg, $n = 9$) on channel conductance. I_{Ca} current traces obtained during 10-ms depolarizations to the reversal potential (V_{rev}) were normalized to the area of the Q_{ON} gating charge of wild-type channels to show the relative increase in tail current. Representative superimposed traces are shown separately for each mutant with wild-type channels (for statistics, see **Supplementary Tables 3** and **4**). (e, f) Effect of $\text{Ca}_v1.3$ alteration p.Gly403Arg. (e) Example of normalized I_{Ca} traces for wild-type and Gly403Arg channels (expressed with $\alpha 2\delta 1$ and $\beta 1b$) elicited by 500-ms depolarization to +10 mV with 10 mM Ca^{2+} charge carrier. (f) Mean percent inactivation at 500 ms was reduced for Gly403Arg mutant channels ($n = 15$ –16) compared to wild-type channels ($n = 14$ –18). Error bars, s.e.m. ** $P = 0.0094$; *** $P < 0.0001$ (Student's unpaired t test).

reports the number of active channels at the plasma membrane), indicating higher average I_{Ca} per active channels (**Fig. 5d** and **Supplementary Table 3**). Similar current properties were seen upon coexpression with palmitoylated $\beta 2a$ subunits, which are known to stabilize slower inactivation kinetics (**Supplementary Tables 3** and **4**). The Gly403Arg mutant was more difficult to analyze because of lower maximum conductance (**Supplementary Fig. 7a**). This was not a result of lower protein expression of this mutant, as shown by immunoblotting of transfected tsA-201 cells (**Supplementary Fig. 7b**). However, Gly403Arg-mediated I_{Ca} showed substantially reduced inactivation over 500 ms, again indicative of gain of function (**Fig. 5e,f**).

Taken together, these gating changes suggest that a zona glomerulosa-like APA harboring these mutant $\text{Ca}_v1.3$ channels would have increased Ca^{2+} entry, causing increased intracellular Ca^{2+} -mediated signaling and, thus, enhanced aldosterone secretion. In a minority of individuals with APAs, an L-type Ca^{2+} channel blocker causes profound reductions in plasma aldosterone concentrations, sometimes masking diagnosis²⁵. Because an L-type Ca^{2+} channel blocker is now a first-line therapeutic option for the treatment of hypertension and because the APAs with *CACNA1D* mutations can be particularly small, they are more likely to be masked or overlooked as a common cause of hypertension.

Because the *CACNA1D* mutations target multiple repetitive sites for gain of function, the seven sites reported here seem unlikely to be

the final tally. In *KCNJ5*, we also identified a previously unreported somatic mutation around the selectivity filter, encoding p.Glu145Lys (**Supplementary Fig. 1**), making this the sixth mutation found in this gene in APAs. In the P-type ATPases, the strongly dominating hotspot seems to be at the cytoplasmic ion entry site, so any additional pump mutations found in APAs would most likely affect this site.

The strong correlation between APAs and recurrent somatic mutations in well-known genes regulating intracellular cations is notable. The frequent coexistence of a zona glomerulosa-like APA with zona fasciculata-like or non-secretory adenoma indicates that the mutations increasing aldosterone synthesis and secretion may be separate from those causing adenoma formation. In our small APAs, usually unchanged in size for many years, synthesis may have superseded proliferation and protected against the increased cell turnover characteristic of low aldosterone production, as seen in the adrenal glands of *Cyp11b2*-null mice and in the thin zona glomerulosa of salt-loaded humans where CYP11B2 expression is often patchy or absent^{10,26,27}. That increased hormone production by APAs is due to constitutively upregulated hormone production, rather than to increased cell mass, was suggested by the increased uptake of the *CYP11B* radiotracer in our positron emission tomography-computed tomography (PET-CT) studies²⁸ (**Fig. 2**) and by the comparison of CYP11B2 mRNA and protein expression levels in APAs and normal zona glomerulosa (**Fig. 2** and **Supplementary Fig. 4**).

In summary, a substantial proportion of APAs resembling adrenal zona glomerulosa cells harbor gain-of-function mutations in genes important for the regulation of Na^+ and Ca^{2+} , *ATP1A1* and *CACNA1D*, respectively. Mutations in both genes appear to be more common in cohorts enriched for smaller, zona glomerulosa-like APAs, highlighting the notion that diagnosis should not rely on finding a definite nodule upon adrenal imaging.

URLs. PyMOL, <http://www.pymol.org/>; OriGene, <http://www.origene.com/>; TreeView software, <http://rana.lbl.gov/EisenSoftware.htm>; Exome Power Calculation software, <http://darth.ssg.uab.edu:8080/epc/index.jsp>.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. Disease-causing variants will be submitted to ClinVar. Exome data are available upon request within a scientific cooperation. Gene expression data are available from the NCBI Gene Expression Omnibus (GEO) under accession [GSE48303](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Cambridge, UK. We are grateful to M. Gurnell for discussion and care of many of the patients, to N. Jamieson for all the laparoscopic adrenalectomies and to A. Marker for pathological diagnosis. We thank Dr. Yasmin for providing peripheral DNA samples from healthy, normotensive subjects. We thank R. Kuc and the Human Research Tissue Bank of Addenbrooke's Hospital, which is supported by the NIHR Cambridge BRC, for help with storage of fresh adrenal tissue for the Cambridge cohort; we particularly acknowledge B. Haynes, D. Walters, K. Brown, M. Elazoui, C. Karpinskyj, M. Bromwich and K. Payne. The work was funded by the British Heart Foundation (PG/07/085/23349), the Wellcome Trust (085686/Z/08/A), the NIHR Cambridge Biomedical Research Centre (Cardiovascular) and an NIHR Senior Investigator award to M.J.B. The work was also supported by the Austin Doyle Award funded by Servier Australia (to E.A.B.A.). C.A.B. is supported by the Wellcome Trust PhD program in Metabolic and Cardiovascular Disease. J.Z. is supported by the Cambridge Overseas Trust and the Sun Hung Kai Properties-Kwoks' Foundation PhD program. G.S.H.Y. was supported by European Union FP7-HEALTH-2009-241592 EurOCHIP and FP7-FOOD-266408 Full4Health. **Aarhus, Denmark.** We thank J. Egebjerg Jensen for discussion of the electrophysiology data. H.P. was supported by grants from The Carlsberg Foundation, The Lundbeck Foundation and L'Oréal/UNESCO. **University of Innsbruck, Austria.** The work was supported by the Austrian Science Fund (F44020). **University College London, UK.** We thank W. Pratt for technical assistance. The work was supported by the Wellcome Trust (098360/Z/12/Z). **Hradec Kralove, Czech Republic.** We thank A. Ryska, who selected the most appropriate adrenal samples for the Czech cohort. Funding is provided by program PRVOUK P037/03. **Nijmegen, The Netherlands.** We thank J.W.M. Lenders for introducing the collaboration and for his leading role in the recruitment of the Dutch cohort.

AUTHOR CONTRIBUTIONS

E.A.B.A. and M.J.B. designed and analyzed the adrenal experiments. H.P. and M.V.C. designed the electrophysiology experiments and performed cloning. H.P. performed and analyzed the electrophysiology experiments. M.V.C. made the homology model, and M.V.C., H.P. and P.N. discussed the structural analyses. A.C.D. and W.M. designed experiments on the Gly403Arg mutant of $\text{Ca}_v1.3$, undertaken by W.M. and K.C. K.C. performed protein blotting. P.T., A.L. and J.S. designed the experiments for the remaining $\text{Ca}_v1.3$ mutants. P.T. cloned the *CACNA1D* mutations, and A.L. performed whole-cell patch-clamp experiments. G.S.H.Y., S.G.N. and I.M. contributed to the design of RNA analyses, including for microarray analysis. E.G.B. and I.S.F. advised on the design and interpretation of exome sequencing. N.R., F.M. and J.H. designed and interpreted microfluidic sequencing. E.A.B.A. performed the H295R transfections with help from J.Z. J.Z. performed genotyping and Sanger sequencing with help from E.A.B.A., C.M., S.G. and E.G.B. Gene expression studies were performed by E.A.B.A., C.A.B., A.E.D.T., J.Z. and L.H.S. W.Z. performed immunohistochemistry, for which A.P.D. designed

selective antisera to CYP11B1 and CYP11B2. For the Cambridge cohort, M.J.B., E.A.B.A., J.Z., C.M. and L.H.S. collected and prepared samples. For the Czech cohort, J.C. and M.S. collected and analyzed the clinical data, and E.A.B.A. and C.M. examined pathology, performed DNA isolation and prepared samples. For the Dutch cohort, J.D. executed the recruitment, B.K. examined pathology, B.T. performed DNA isolation, and T.D. prepared the samples. E.A.B.A. prepared the supplementary information, and I.S.F., J.S., H.P. and M.J.B. wrote the manuscript with comments from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Choi, M. *et al.* K^+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. *Science* **331**, 768–772 (2011).
- Beuschlein, F. *et al.* Somatic mutations in *ATP1A1* and *ATP2B3* lead to aldosterone-producing adenomas and secondary hypertension. *Nat. Genet.* **45**, 440–444 (2013).
- Azizan, E.A. *et al.* Microarray, qPCR and *KCNJ5* sequencing of aldosterone-producing adenomas reveal differences in genotype and phenotype between zona glomerulosa- and zona fasciculata-like tumors. *J. Clin. Endocrinol. Metab.* **97**, E819–E829 (2012).
- Rossi, G.P. *et al.* A prospective study of the prevalence of primary aldosteronism in 1,125 hypertensive patients. *J. Am. Coll. Cardiol.* **48**, 2293–2300 (2006).
- Rossi, G.P. A comprehensive review of the clinical aspects of primary aldosteronism. *Nat. Rev. Endocrinol.* **7**, 485–495 (2011).
- Boukroun, S. *et al.* Prevalence, clinical, and molecular correlates of *KCNJ5* mutations in primary aldosteronism. *Hypertension* **59**, 592–598 (2012).
- Funder, J.W. The genetics of primary aldosteronism: chapter two. *Hypertension* **59**, 537–538 (2012).
- Mazzocchi, G. *et al.* Ghrelin enhances the growth of cultured human adrenal zona glomerulosa cells by exerting MAPK-mediated proliferogenic and antiapoptotic effects. *Peptides* **25**, 1269–1277 (2004).
- Shigematsu, K. *et al.* Primary aldosteronism with aldosterone-producing adenoma consisting of pure zona glomerulosa-type cells in a pregnant woman. *Endocr. Pathol.* **20**, 66–72 (2009).
- Wolkersdorfer, G.W. *et al.* Differential regulation of apoptosis in the normal human adrenal gland. *J. Clin. Endocrinol. Metab.* **81**, 4129–4136 (1996).
- Morth, J.P. *et al.* Crystal structure of the sodium-potassium pump. *Nature* **450**, 1043–1049 (2007).
- Winther, A.M. *et al.* The sarcolipin-bound calcium pump stabilizes calcium sites exposed to the cytoplasm. *Nature* **495**, 265–269 (2013).
- Toyoshima, C. *et al.* Crystal structures of the calcium pump and sarcolipin in the Mg^{2+} -bound E1 state. *Nature* **495**, 260–264 (2013).
- Li, C., Capendeguy, O., Geering, K. & Horisberger, J.D. A third Na^+ -binding site in the sodium pump. *Proc. Natl. Acad. Sci. USA* **102**, 12706–12711 (2005).
- Poulsen, H. *et al.* Neurological disease mutations compromise a C-terminal ion pathway in the Na^+/K^+ -ATPase. *Nature* **467**, 99–102 (2010).
- Vasilyev, A., Khater, K. & Rakowski, R.F. Effect of extracellular pH on presteady-state and steady-state current mediated by the Na^+/K^+ pump. *J. Membr. Biol.* **198**, 65–76 (2004).
- Efthymiadis, A., Rettinger, J. & Schwarz, W. Inward-directed current generated by the Na^+/K^+ pump in Na^+ - and K^+ -free medium. *Cell Biol. Int.* **17**, 1107–1116 (1993).
- Brale, L.M. & Williams, G.H. The effects of ouabain on steroid production by rat adrenal cells stimulated by angiotensin II, $\alpha 1$ -24 adrenocorticotropin, and potassium. *Endocrinology* **103**, 1997–2005 (1978).
- Yingst, D.R., Davis, J., Krenz, S. & Schiebinger, R.J. Insights into the mechanism by which inhibition of Na^+/K^+ -ATPase stimulates aldosterone production. *Metabolism* **48**, 1167–1171 (1999).
- Cushman, P. Jr. Inhibition of aldosterone secretion by ouabain in dog adrenal cortical tissue. *Endocrinology* **84**, 808–813 (1969).
- Kau, M.M., Kan, S.F., Wang, J.R. & Wang, P.S. Inhibitory effects of digoxin and ouabain on aldosterone synthesis in human adrenocortical NCI-H295 cells. *J. Cell Physiol.* **205**, 393–401 (2005).
- Hemara-Wahanui, A. *et al.* A *CACNA1F* mutation identified in an X-linked retinal disorder shifts the voltage dependence of $\text{Ca}_v1.4$ channel activation. *Proc. Natl. Acad. Sci. USA* **102**, 7553–7558 (2005).
- Splawski, I. *et al.* Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl. Acad. Sci. USA* **102**, 8089–8096 (2005).
- Hoda, J.C., Zaghetto, F., Koschak, A. & Striessnig, J. Congenital stationary night blindness type 2 mutations S229P, G369D, L1068P, and W1440X alter channel gating or functional expression of $\text{Ca}_v1.4$ L-type Ca^{2+} channels. *J. Neurosci.* **25**, 252–259 (2005).
- Brown, M.J. & Hopper, R.V. Calcium-channel blockade can mask the diagnosis of Conn's syndrome. *Postgrad. Med. J.* **75**, 235–236 (1999).
- Lee, G. *et al.* Homeostatic responses in the adrenal cortex to the absence of aldosterone in mice. *Endocrinology* **146**, 2650–2656 (2005).
- Nishimoto, K. *et al.* Adrenocortical zonation in humans under normal and pathological conditions. *J. Clin. Endocrinol. Metab.* **95**, 2296–2305 (2010).
- Burton, T.J. *et al.* Evaluation of the sensitivity and specificity of ^{11}C -metomidate positron emission tomography (PET)-CT for lateralizing aldosterone secretion by Conn's adenomas. *J. Clin. Endocrinol. Metab.* **97**, 100–109 (2012).

ONLINE METHODS

Subjects. Individuals with unilateral primary aldosteronism were recruited from three centers (Addenbrooke's Hospital, University of Cambridge, $n = 63$; University Hospital Hradec Kralove, $n = 39$; Radboud University Nijmegen Medical Centre, $n = 50$). Case detection and subtype identification were in accordance with institutional guidelines. Primary hyperaldosteronism was diagnosed by an elevated aldosterone/renin ratio (ARR) and followed up by confirmatory studies (CT scan or MRI and/or adrenal venous sampling and/or [^{11}C]-metomidate PET-CT scan²⁸). Adrenalectomy reversed the biochemical abnormalities. Cambridge and Czech subjects gave written informed consent for genetic investigation, which was approved by the Cambridgeshire 2 Research Ethics Committee and the University Hospital Hradec Kralove Ethics Committee, respectively. Samples from Dutch subjects were used in accordance with the code of conduct of research with human material in The Netherlands.

For exome sequencing, we selected ten zona glomerulosa-like APAs ($\geq 50\%$ compact zona glomerulosa-like cells and low expression of *CYP17A1*) and three zona fasciculata-like APAs with somatic *KCNJ5* mutations (as sensitivity controls) from the Cambridge cohort. Paired genomic DNA for 9 of the 10 zona glomerulosa-like APAs from either venous blood (subject 8) or the peritumoral tissue (subjects 1–5, 7 and 9) were also exome sequenced.

We estimated sample size in two ways. The cruder estimation showed that if 30% of samples had a new mutation in the same gene, there would be $(0.7)^9$ ($=0.04$) probability of not finding at least two mutated samples in a cohort of ten. The public domain software of Zhi and Chen (Exome Power Calculation software) showed that, in ten affected individuals, a 40% mutation frequency would be required to provide 80% power, if sequencing sensitivity were 99% and ten genes passed the somatic mutation filter.

Nucleic acid extraction. DNA or RNA was extracted from 209 adenomas/nodules in 152 affected individuals. In addition, DNA or RNA was extracted from 51 paired peritumoral adrenal cortices and 3 paired peripheral DNA samples, 11 non-aldosterone-secreting adenomas (8 Cushing adenomas, 2 adrenal mass and 1 incidentaloma) and 100 peripheral DNA samples from healthy, normotensive subjects. DNA was extracted using standard procedures. Total DNA-free RNA was isolated using the TRIzol Plus RNA Purification System (12183-555, Life Technologies) along with the PureLink DNase Set (12185010, Life Technologies). Reverse transcription of 1 μg of RNA was performed using the Reverse Transcriptase System (Promega) with a 1:1 mixture of random hexamer and oligo(dT) primers according to the manufacturer's instructions.

Exome sequencing. Exome sequencing was performed by BGI Shenzhen. In brief, qualified genomic DNA was randomly fragmented by Covaris, resulting in DNA fragments with a base-pair peak at 200 to 300 bp, and adaptors were then ligated to the resulting fragments' ends. Extracted DNA was amplified by ligation-mediated PCR, purified and hybridized to the NimbleGen SeqCap EZ Exome 44M array (Roche NimbleGen) for enrichment. Each captured library was subjected to high-throughput sequencing using the HiSeq2000 platform, and all samples achieved $>30\times$ read coverage. Raw image files were processed by Illumina base calling Software 1.7 with default parameters. Sequences for each library were generated as 90-bp paired-end reads.

Variant detection. After removing reads containing sequencing adaptors and low-quality reads, high-quality single-end reads were aligned to human genome Build 37(hg19) using Burrows-Wheeler Aligner (BWA). Potential SNPs were detected by SOAPsnp, and potential small insertion-deletions (indels) were detected by SAMtools. Standard quality control was performed at each stage of the analysis pipeline for the clean data, the alignment and the called variant.

Pairwise comparison analysis of single-nucleotide variants (SNVs), somatic indels and CNVs were detected by VarScan, GATK and ExomeCNV, respectively. In the pairwise comparison analysis, several heuristic rules were applied: (i) both the tumor and matched normal samples should be covered sufficiently ($\geq 10\times$) at the genomic position compared; (ii) the average base quality for a given genomic position should be at least 15 in both the tumor and normal samples; (iii) the variants should be supported by at least 10% of the total reads

in the tumors, and no reads supporting high-quality variants were allowed in normal controls; and (iv) the variants should be supported by at least five reads in the tumors. Then, ANNOVAR was used to annotate the variant results, with variants passing the following three quality checks defined as high-probability somatic mutations: (i) the somatic mutation had a high probability of occurring only in the APA ($P \leq 1 \times 10^{-4}$); (ii) the somatic mutation did not exhibit strand bias, where depth of the supporting reads on one strand was less than $5\times$; and (iii) depth of the supporting reads at that locus was at most $200\times$.

Variant sequencing. Confirmation of 23 of the 24 high-probability somatic mutations that passed the quality check filters on exome sequencing was performed by PCR amplification and Sanger sequencing. Mutations in *ATP1A1*, *CACNA1D* or *KCNJ5* detected by exome sequencing, TaqMan genotyping or microfluidic sequencing were confirmed by PCR amplification and Sanger sequencing.

ATP1A1 and CACNA1D genotyping. DNA or cDNA of Cambridge samples and DNA of Dutch and Czech samples were genotyped using custom TaqMan genotyping assays (Applied Biosystems) for the substitution mutations found in *ATP1A1* (encoding p.Leu104Arg) and *CACNA1D* (encoding p.Val259Asp, p.Gly403Arg, p.Pro1336Arg and p.Ile1750Met).

Microfluidic sequencing. Multiplex target-specific amplification was performed on the Access Array microfluidic system (Fluidigm) according to the manufacturer's recommendations. Target-specific primers (TS-Forward or TS-Reverse) were designed with a custom pipeline to tile the coding regions of genes *ATP1A1* (Ensembl transcript ENST00000295598), *CACNA1D* (ENST00000288139) and *KCNJ5* (ENST00000529694). Amplicon lengths varied between 151 and 395 bp (average of 220 bp); melting temperatures varied between 57.4°C and 61.2°C (average of 59.9°C). Each target-specific primer consists of a universal 5' end and a target-specific 3' end.

After PCR products were barcoded using a 10-base indexing system, they were analyzed using Agilent 2100 BioAnalyzer (Supplementary Fig. 8). Single-end sequencing of 150 bases of pooled library was performed on an Illumina Genome Analyzer IIx sequencer using custom sequencing primers targeted to the CS1 and CS2 tags according to the manufacturer's recommendations.

Microarray analysis. RNA was analyzed on GeneChip Human Gene 1.0 ST arrays (Affymetrix). Each array comprises 764,885 distinct probes, which interrogate 28,869 well-annotated genes (Affymetrix). Following hybridization, arrays were washed and stained with a streptavidin-phycoerythrin conjugate using an automated protocol on a GeneChip Fluidics Station 450 followed by scanning on a GeneChip (GCS3000) Scanner. Quality control, data processing and analysis were performed using GeneChip Command Console Software (Affymetrix) and Partek Genomic Suite v. 6.5. All samples were normalized by the GC-RMA method (gene chips-robust multichip analysis) with quantile normalization and median polish for probe set summarization. False discovery rate control (Benjamini-Hochberg method) was used to correct for multiple testing. Differentially expressed genes were defined by a fold-change difference of >2 and $P < 0.05$.

Plasmids. Plasmids encoding human $\alpha 1$ and $\beta 1$ subunits of the Na^+/K^+ ATPase were purchased from OriGene and subcloned into the pXOON vector using EcoRI and NotI. Mutations encoding p.Gln118Arg and p.Asn129Asp were introduced into $\alpha 1$ by PCR to reduce ouabain sensitivity, yielding the construct referred to as wild type. The other mutations and deletions were also introduced via PCR.

The human wild-type $\text{Ca}_v1.3$ channel (*CACNA1D* gene, GenBank accession EU363339) containing alternative exons 8a and 42 was previously cloned into pGFP⁺ vector (no GFP tag and mammalian system expression controlled by CMV promoter)²⁹. The mutations encoding the p.Val259Asp, p.Gly403Arg, p.Ile1750Met and p.Pro1336Arg alterations were cloned in the above mentioned construct using a standard PCR approach.

Na^+/K^+ ATPase electrophysiology. RNA was transcribed from NheI-digested plasmids with the mMESSAGE ULTRA kit (Ambion). RNA for $\beta 1$ (1 ng) and $\alpha 1$ (10 ng) were coinjected into oocytes from *X. laevis*. After 1–6 d at

15 °C, two-electrode voltage clamping was performed in 115 mM Na, 110 mM sulfamic acid, 1 mM MgCl₂, 0.5 mM CaCl₂, 5 mM BaCl₂, 10 mM HEPES and 1 μM ouabain. Unless otherwise indicated, the pH was 7.4. In potassium-containing buffers, sodium was replaced by equimolar potassium, and in sodium-free solutions, NMDG replaced sodium. To determine steady-state currents, a series of 200-ms 20 mV steps was run, and the 10 mM ouabain background current was subtracted.

Ca_v1.3 electrophysiology. Cell culture and transient expression of Ca_v1.3 constructs in tsA-201 cells were performed as described³⁰. Whole-cell patch-clamp recordings were performed at room temperature. Borosilicate glass electrodes were pulled (micropipette puller, Sutter instruments) and fire polished (microforge, Narishinge MF-830) at a final resistance of 1.5–2.5 M. Cells were recorded at a sampling rate of 2–5 kHz using an Axopatch 200B amplifier (Axon Instruments), digitized with Digitizer 1322A (Axon Instruments) and recorded with pClamp 10.2 software (Axon Instruments). Recording solution consisted of the following: bath solution, 15 mM or 2 mM CaCl₂, 10 mM HEPES, 150 mM or 170 mM choline-Cl and 1 mM MgCl₂, adjusted to pH 7.4 with CsOH; intracellular solution, 135 mM CsCl, 10 mM HEPES, 10 mM Cs-EGTA and 1 mM MgCl₂ adjusted to pH 7.4 with CsOH. Cells were held at a holding potential of –80 mV before a step protocol of 10 ms to different voltages was applied to determine the current-voltage relationship. Currents were leak subtracted using a P/4 protocol. Inactivation time course was measured during 5-s depolarizations from –80 mV to V_{\max} and fitted to a standard double-exponential decay using GraphPad Prism 5 (GraphPad Software). The voltage dependence of inactivation was measured by applying a control test pulse (10 ms to V_{\max}) followed by a 5-s conditioning step and a subsequent 20-ms test pulse to V_{\max} (30-s recovery between protocols). Inactivation was calculated as the ratio between the current amplitudes of the test versus control pulse. As an estimate of the changes in single-channel properties, the ionic tail current during repolarization following a 10-ms depolarization step pulse to the reversal potential was normalized to the ON-gating current obtained upon depolarization in the same pulse (**Supplementary Table 3**). Current-voltage curves were fitted to the equation $I = G_{\max}(V - V_{\text{rev}})/(1 + \exp^{-(V - V_{0.5})/k})$, where V_{rev} is the reversal potential, V the test potential, I the peak current, G_{\max} the maximum conductance, $V_{0.5}$ the half-maximal activation voltage and k the slope. The voltage dependence of Ca²⁺ conductance was fitted according to a Boltzman distribution $G = G_{\max}/(1 + \exp^{-(V - V_{0.5})/k})$. Steady-state inactivation parameters were obtained by fitting the data to a modified Boltzmann equation $G = (1 - G_{\max})/(1 + \exp^{(V - V_{0.5})/k}) + G_{\max}$. To reduce noise in some experiments, protocols were repeated up to five times, and recordings were averaged as described³¹.

tsA-201 cell lysis and immunoblotting. After 72 h of transfection with vectors encoding wild-type and mutant hCa_v1.3 and β1b, tsA-201 cells were homogenized in PBS, pH 7.4 at 4 °C containing 1% IGEPAL and protease inhibitors (Complete, Roche), were sonicated for 10 s and were incubated on ice for 45 min. Whole-cell lysates were then centrifuged at 14,000g for 30 min at 4 °C, and pellets were discarded. Aliquots of supernatant were assayed for total protein (Bradford assay; Bio-Rad). Aliquots of whole-cell lysate corresponding to 20 μg of total protein were diluted with Laemmli sample buffer supplemented with 100 mM DTT and 25 mM N-ethylmaleimide and were resolved by SDS-PAGE on 3–8% Tris-acetate gels (Invitrogen) and transferred to polyvinylidene fluoride membrane (Bio-Rad) by protein blotting (semi-dry; Bio-Rad). The primary antibodies used included antibody to Ca_v1.3 (mouse monoclonal, Neuromab; 3 μg/ml) and antibody to GAPDH (mouse monoclonal; Ambion, MAB 5718; 1:25,000 dilution), and the secondary antibody was HRP-conjugated goat antibody to mouse (Bio-Rad). Signal was obtained by HRP reaction with fluorescent product (ECL Plus, GE Healthcare) and membranes were scanned on a Typhoon 9410 phosphorimager (GE Healthcare).

Cell culture and experimentation. H295R human adrenocortical carcinoma cells (originally obtained from ATCC) were cultured in DMEM/Nutrient

F-12 Ham supplemented with 10% FCS, 100 U of penicillin, 0.1 mg/ml streptomycin, 0.4 mM L-glutamine and ITS (insulin-transferrin-sodium selenite medium) at 37 °C in 5% CO₂. Cells were transfected with either vector control or ouabain-resistance vector encoding wild-type ATP1A1 or the Leu104Arg ATP1A1 mutant using standard procedures. Transfected cells were plated into 24-well plates (100,000 cells per well) in 0.5 ml of growth medium. After 48 h of transfection, H295R cells were serum deprived in unsupplemented medium for 24 h and were then incubated in fresh medium with the treatments specified in the legend to **Figure 2**. Supernatants for aldosterone concentration measurement were collected after respective incubation time of treatments, and cells were harvested for analysis of mRNA. Genotypes of the transfected cells were confirmed by TaqMan genotyping.

Aldosterone concentration measurements. Aldosterone concentrations were determined by ¹²⁵I radioimmunoassay using a commercially available Coat-A-Count kit (Diagnostic Products Corp). Aldosterone concentrations were normalized to total cell protein, determined by extraction of protein with lysis buffer and BCA protein assay (Pierce Biotechnology).

Laser capture microdissection (LCM). LCM was used to acquire samples of zona fasciculata and zona glomerulosa cells in peritumoral adrenal tissue from eight individuals with pheochromocytoma. Procedures for sample acquisition and methodology have been described in detail elsewhere⁵.

Quantification of mRNA expression. Cells were kept in RNeasy lysis buffer until RNA was extracted. Total RNA was isolated and reverse transcribed using methods mentioned previously. mRNA expression of genes of interest was quantified using commercially available TaqMan ABI probes (Applied Biosystems), and *CYP11B2* and *CYP11B1* expression was quantified using custom-made TaqMan probes (Invitrogen) that had been validated for specificity⁵. The housekeeping 18S rRNA (Applied Biosystems) was used for normalization.

Immunohistochemistry. Immunohistochemistry was performed on formalin-fixed, paraffin-embedded adrenal sections (4 μm) using an automated immunostainer with cover tile technology (Bond-III system, Leica Biosystems). Commercial antibodies to ATP1A1 (A276, Sigma; 1:1,000 dilution), *CACNA1D* (clone N38/8, UC Davis/NIH NeuroMab Facility; 1:500 dilution) and *KCNJ5* (HPA017353, Sigma; 1:100 dilution) and custom-made antibodies to *CYP11B2* (1:10 dilution) and *CYP11B1* (1:100 dilution) were used as the primary antibodies. The antiserum selective for *CYP11B1* was affinity purified against the immunizing antigen and was subsequently purified against a *CYP11B2* antigen column to remove any cross-reacting antibodies. Selective antisera were generated in the same way for *CYP11B2*. Negative control experiments, in which primary antibodies were omitted, resulted in a complete absence of staining. Images were captured using a standard bright-field microscope, a U-TV1-X digital camera and CellD software (Olympus UK).

Statistical analysis. If not stated otherwise, group results are expressed as mean values with s.e.m. and compared using Student's unpaired *t* tests. The significance level of $P < 0.05$ was considered to indicate statistical significance. Statistical analysis was performed using standard statistical software. For *ATP1A1* electrophysiological data, statistical analysis was performed by individual one-way ANOVA with the Bonferroni *post-hoc* test. For *CACNA1D* electrophysiological data, statistical analysis was performed by Clampfit 10.2 (Axon Instruments) and Sigma Plot 12 (Systat Software).

29. Koschak, A. *et al.* α1D (Ca_v1.3) subunits can form L-type Ca²⁺ channels activating at negative voltages. *J. Biol. Chem.* **276**, 22100–22106 (2001).
30. Singh, A. *et al.* C-terminal modulator controls Ca²⁺-dependent gating of Ca_v1.4 L-type Ca²⁺ channels. *Nat. Neurosci.* **9**, 1108–1116 (2006).
31. Baig, S.M. *et al.* Loss of Ca_v1.3 (*CACNA1D*) function in a human channelopathy with bradycardia and congenital deafness. *Nat. Neurosci.* **14**, 77–84 (2011).



Multi-genome alignment for quality control and contamination screening of next-generation sequencing data

James Hadfield^{1*} and Matthew D. Eldridge^{2*}

¹ Genomics Core Facility, Cancer Research UK Cambridge institute, University of Cambridge, Cambridge, UK

² Bioinformatics Core Facility, Cancer Research UK Cambridge institute, University of Cambridge, Cambridge, UK

Edited by:

Mick Watson, The Roslin Institute, UK

Reviewed by:

Jianlin Cheng, University of Missouri, Columbia, USA

Stephen Taylor, Weatherall Institute of Molecular Medicine, UK

*Correspondence:

James Hadfield, Genomics Core Facility, Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
e-mail: james.hadfield@cruk.cam.ac.uk;

Matthew D. Eldridge, Bioinformatics Core Facility, Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
e-mail: matthew.eldridge@cruk.cam.ac.uk

[†]James Hadfield and Matthew D. Eldridge have contributed equally to this work.

The availability of massive amounts of DNA sequence data, from 1000s of genomes even in a single project has had a huge impact on our understanding of biology, but also creates several problems for biologists carrying out those experiments. Bioinformatic analysis of sequence data is perhaps the most obvious challenge but upstream of this even basic quality control of sequence run performance is challenging for many users given the volume of data. Users need to be able to assess run quality efficiently so that only high-quality data are passed through to computationally-, financially-, and time-intensive processes. There is a clear need to make human review of sequence data as efficient as possible. The multi-genome alignment tool presented here presents next-generation sequencing run data in visual and tabular formats simplifying assessment of run yield and quality, as well as presenting some sample-based quality metrics and screening for contamination from adapter sequences and species other than the one being sequenced.

Keywords: next-generation sequencing, quality control, contamination screen

INTRODUCTION

It is vital in any laboratory to assess the quality of the data being generated. In a next-generation sequencing (NGS) facility the volumes of data can be overwhelming and automated quality control (QC) reporting is an ideal. There are many metrics to consider when looking at a sequencing run, some are run specific, others sample specific and many can be affected by both run and sample. Understanding what individual metrics mean in a particular context is complex and can require significant experience. Tools that help simplify analysis by building on this experience and removing subjectivity are becoming increasingly vital. We have developed the multi-genome alignment (MGA) contamination screen that can be used to calculate a few key, simple but important metrics, primarily data yield and quality, whilst also providing some additional sample related QC.

Tools like MGA are not new. Perhaps the earliest example of a QC tool is the Phred (Ewing et al., 1998) package developed to improve methods for gel-based Sanger-sequencing trace analysis and base quality scoring. The abstract of this paper written 15 years ago is surprisingly relevant today, stating: “it is particularly important that human involvement in sequence data processing be significantly reduced or eliminated” and that there is a need to “make human review (of sequence data) more efficient.” Almost

every molecular biologist working today has seen and benefitted from this work in their Sanger sequencing results. Life Technologies (formerly Applied Biosystems, Santa Clara, CA, USA) produced a free tool, Sequence Scanner v1.0 (Applied Biosystems, 2005), that allowed a very quick visual check of 96 samples. This move away from inspection of individual traces to a more gross assessment of a Sanger sequencing run was necessitated by the increase in sample volumes due to the introduction of automated capillary sequencers. The introduction of microarrays was accompanied by QC tools that allowed the vast amounts of data to be assessed before starting complex analysis pipelines. Two of the major providers included such tools; Affymetrix (Santa Clara, CA, USA) provided a simple text-based reporting tool in their MAS5.0 (Hubbell et al., 2002) primary data analysis package and Agilent Technologies (San Francisco, CA USA) provided a very comprehensive visual, graphical, and text-based tool in their Feature Extraction software (Agilent, 2013). For NGS data the most widely used software, not provided by instrument vendors, for quality analysis of data is FastQC (Andrews, 2010), which presents multiple metrics for each dataset, including per base sequence quality score, per base GC content and duplication rate. A feature of the most of the tools above is their reliance on multiple metrics to report on the quality of what are complex assays. However, the

use of individual metrics must be evaluated carefully alongside others including the starting sample QCs, and in the context of the sample or run being considered.

The MGA tool presented here aims to provide a subset of metrics that can be quickly assessed with minimal explanation, and allow users of NGS data to determine if the data generated are of sufficient yield and quality. The tool is not intended to be comprehensive nor used in isolation, rather as part of a formal assessment of experimental quality.

MATERIALS AND METHODS

The MGA contaminant screen is an alignment-based method for detecting contamination in genomic or transcriptomic sequencing libraries. A sample of read sequences and base quality scores is extracted from the FASTQ files produced by the sequencing instrument. In practice, we have found that a sample of 100,000 reads is sufficient to detect moderately low levels of contamination. This represents a small fraction of the data usually generated from a lane of sequencing. The screen can be run on any number of FASTQ datasets so it would be feasible to look separately at every library in a multiplexed pool to determine the likely contamination in each. We generally assess contamination at the lane level on the basis that libraries from different sources are not typically grouped together on the same lane.

Two differing alignment approaches are taken for (1) identifying sequences likely to have originated from a different species to that being sequenced and (2) detecting the presence of adapter sequences ligated to the ends of sequence fragments. The screen is not capable of detecting contaminant sequences from the same species as that being sequenced.

For detecting cross-species contamination the sampled reads are trimmed to 36 bases and aligned using bowtie (Langmead et al., 2009) to a set of reference genome sequences representing possible contaminants. This includes several mammalian species that are used in our laboratory as well as several thousand bacteria, viruses, and other microorganisms. The latter are grouped together so that, for example, the sampled reads are aligned to a collection of bacterial reference genome sequences and results reported for the set; consequently, the screen may detect bacterial contamination but will not be specific about the contaminant species. We choose to trim the read sequences so that the results from different sequencing runs can be compared and to derive baseline alignment and error (or mismatch) rates; trimming also helps reduce the computational cost and allows for detection of contaminants in runs with adapter contamination (see below). The reads are also aligned to the reference sequence for phi X 174 bacteriophage, commonly used as a spike-in control. Controls are differentiated from target species and contaminants in the final report.

The alignment results for each species, or collections of reference genomes in the case of bacteria, viruses, and fungi, are collated and species are ranked based on the number of reads aligning to each. Each read may align to multiple species as a result of sequence homology between species. To distinguish likely contamination from sequence homology each read is assigned to a single species based on the above ranking. For example, if the target species is human, some of the reads may also align to the mouse genome. Assuming that more reads align to the human

reference sequence than that for mouse, all reads that align to both will be assigned to human, and only those that uniquely map to the mouse genome (and not another higher-ranked species) will be assigned to mouse.

The method for detecting sequencing adapters differs because it is possible that only a part of a read sequence is adapter. This can occur when the genomic fragment is shorter than the number of bases sequenced such that the sequencing runs through to adapter on the 3' end. Accordingly, we report adapter contamination separately since sequences can be associated both with cross-species and adapter contamination. The sampled reads are first converted to FASTA format and then aligned to a set of adapter and primer sequences using the exonerate sequence alignment tool (Slater and Birney, 2005); this is run using a local alignment model with affine gaps, similar to the Smith–Waterman–Gotoh algorithm (Smith and Waterman, 1981; Gotoh, 1982).

Results are presented in both a tabular and graphical form (Figure 1 and supplementary file), the latter as a stacked bar chart in which each portion of the bar represents the assigned reads for a particular species. The bars are colored green if they match the target species, orange if they match the control, and red if they match another species. The transparency of the bars is adjusted depending on the error or mismatch rate of alignments for the species, with lower mismatch rates corresponding to more opaque bars drawing attention to likely contamination. Adapter contamination is displayed as a separate mauve bar.

The various sampling, trimming, conversion, alignment, and collation steps are defined in an analysis workflow and executed using an in-house workflow management system on a high-performance compute cluster. The MGA screen can also be run on a multi-processor server or high-end workstation. For a single dataset or lane, we align to reference sequences for 23 species and collections of several 1000 bacteria, viruses, and fungi. Each alignment job takes approximately 5 min and the results for eight lanes of an Illumina HiSeq-2000 flow cell are usually available within 15 min of the FASTQ sequence data being available on the compute cluster (overall CPU is around 3–4 hours).

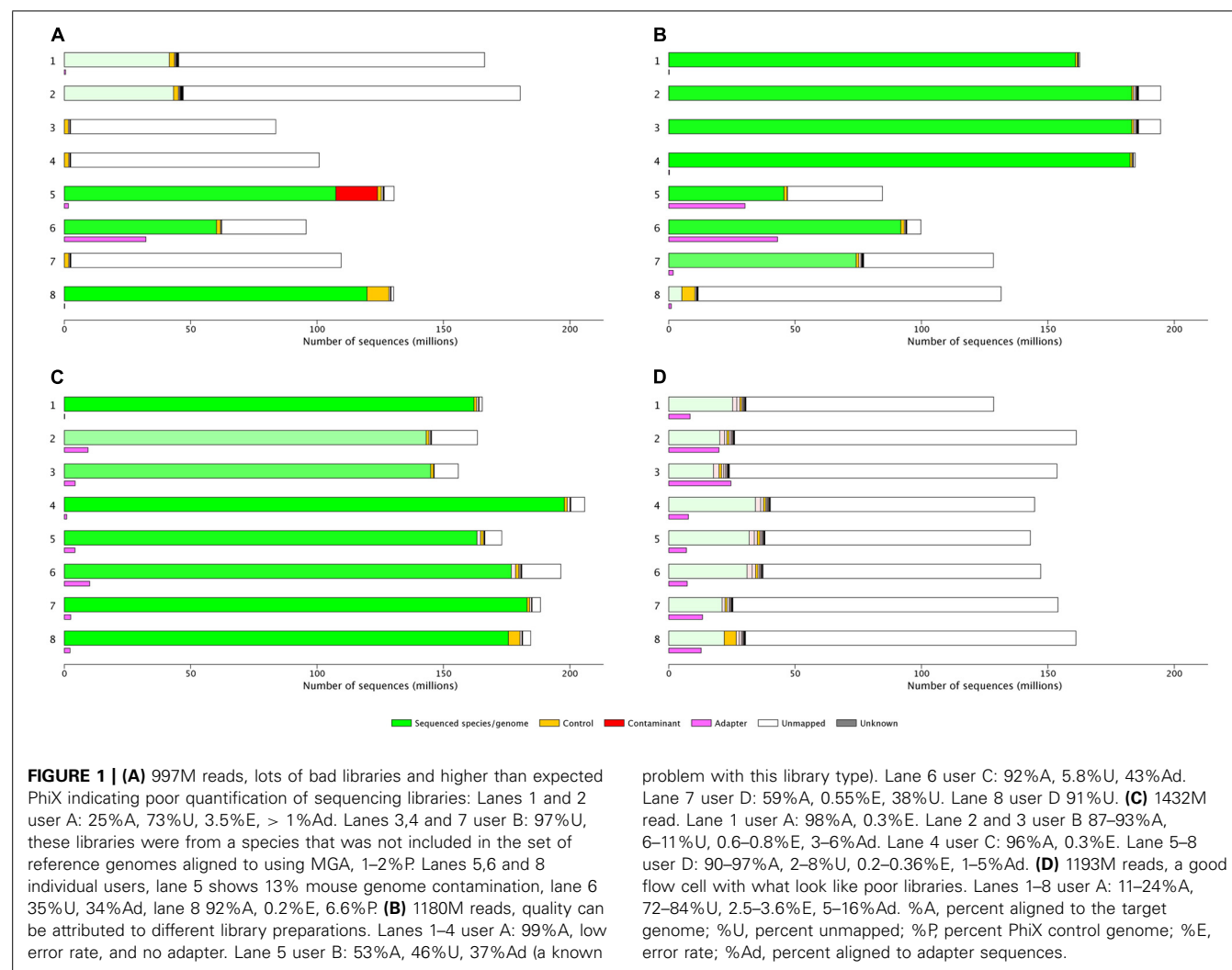
The software, as well as instructions for installing and running MGA, are available here: <https://github.com/crukci-bioinformatics/MGA>

RESULTS

The MGA tool has been used for every sequencing run performed at the Cambridge Institute genomics core for the past three and a half years, on Illumina's (San Diego, CA, USA) GAIIX, HiSeq, and MiSeq platforms.

This graphical representation allows very quick estimation of the yield and quality of each flow cell or lane.

It is relatively simple to determine the difference between “good” and “bad” flow cells (Figures 1C,D), “good” and “bad” samples (Figure 1B lanes 1–4 vs lanes 5 and 6 or 7 or 8) or flow cells which will require significant investigation by the sequencing lab (Figure 1A) or the user (Figure 1D). However, the interpretation of results needs to be taken in context of the type of library or run, as either can significantly influence QC metrics. The flow cell



shown in **Figure 1D** is known to be a reduced representation bisulfite sequencing (Meissner et al., 2005) run and the alignment of this data is expected to be poor; the high yield of this flow cell suggests the user will be happy with the results generated and no further investigation is likely to be necessary. Flow cell B (**Figure 1B**) lanes 5 and 6 show high adapter contamination of the sequencing lane; this is likely to point to issues with sample preparation in the laboratory where the samples originated. These examples demonstrate how MGA can facilitate sequencing users identification of issues with particular sequencing lanes/flow cells.

DISCUSSION

The MGA contaminant-screen tool was originally conceived to answer queries about contamination in the sequencing process. Contamination can occur at any point along the sequencing process, in a research laboratory where samples are being extracted and libraries prepared, or in a sequencing facility where many thousands of libraries are being handled. An early analysis script simply interrogated the level of PhiX in each lane as we hypothesized that if contamination arose in the sequencing laboratory then PhiX, which should only appear in lane 8 (the control lane) would

also be present in lanes 1–7 at variable levels. Analysis confirmed that significant PhiX contamination in lanes 1–7 was limited to a handful of flow cells.

The utility of the tool in this instance demonstrated how useful a similar approach would be as part of our routine QC of each sequencing lane. The use of a control lane increases sequencing costs by 12.5% and is no longer routine. We moved to a process of unbalanced loading of PhiX: 1% in lanes 1–7 and 5% in lane 8. This simple method allows us to detect any inversion of the flow cell, and to determine if low yield is the result of a sequencing or sample issue. If low yield is due to poor clustering/sequencing then the percentage of PhiX will be as expected, whereas if it is due to poor library quantification then the percentage of PhiX will be incorrect, in any low-yield lanes. This has become an important tool in deciding how and when to repeat sequencing runs with low yield, and determining who should pay for the repeat lane(s).

When designing the MGA visualization we considered the metrics most useful to determine the yield and quality of a particular sequencing run. As an Illumina run can contain one or two flow cells, and as most flow cell lanes contain a single sample, we present results in a per lane format. We also tried to consider

the context that these reports might be used in and the limitations our methods might have. Illumina provide many QC metrics in their instrument control software. Commonly analyzed metrics are yield, percentage passed filters (%PF), error rate, phasing and prephasing, cluster density, and per cycle reporting of Q-score and percent Q30 data. The very commonly used FastQC tool provides a modular set of analyses that imports data from BAM, SAM, or FASTQ files and generates eleven summary plots including basic run statistics including number of reads, per base sequence quality, and duplicate sequences. A more recent tool is Illumina's QC app (Illumina Inc, 2013), which generates an automated Library QC report containing several QC metrics in tabular and visual form. The MiSeq QC app incorporates a diversity estimate (Daley and Smith, 2013) for each sample that can be used to determine the limit of sequencing depth. The MGA primarily visualizes two details important to all users and managers of NGS data; yield and quality, it also presents data that can be useful in determining why a particular run/lane is sub-optimal in the accompanying tables.

Multi-genome alignment is one tool that core facility managers, bioinformaticians or users can use to assess their sequence data. The use of multiple tools can be confusing so in most cases users will limit themselves to one or two methods. However, there is not currently a single QC tool for NGS data that provides all the metrics users might require, and different types of user will require different tools at different times. MGA allows very quick interpretation of per lane yield and quality with minimal explanation, allowing the Genomics Core facility at the Cancer Research UK Cambridge Institute to inspect each of approximately 2000 lanes per year.

ACKNOWLEDGMENTS

The MGA tool has been developed at the Cambridge Institute over the past four years: we thank Sarah Aldridge, Gordon D. Brown, Kevin Howe, Nik Matthews, and Rory Stark for useful discussions, Anne Pajon and Richard Bowers for help with implementation and maintenance of the MGA tool in our sequencing pipelines, and Cancer Research UK and the University of Cambridge for funding the Bioinformatics and Genomics core facilities through the Cambridge Institute grant.

REFERENCES

Agilent. (2013). *Feature Extraction Software Product Number: G4460AA*. Available at: <http://www.genomics.agilent.com/en/Microarray-Scanner-Processing->

[Software/Feature-Extraction-Software/?cid=AG-PT-144&tabId=AG-PR-1050](http://www.genomics.agilent.com/en/Microarray-Scanner-Processing-Software/Feature-Extraction-Software/?cid=AG-PT-144&tabId=AG-PR-1050)

Andrews, S. (2010). *Fastqc A Quality Control Tool For High Throughput Sequence Data*. Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
Applied Biosystems. (2005). *Sequence Scanner v1.0*. Available at: <https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=600583&tab=Overview>

Daley, T., and Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nat. Methods* 10, 325–327. doi: 10.1038/nmeth.2375

Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185 doi: 10.1101/gr.8.3.175

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162, 705–708. doi: 10.1016/0022-2836(82)90398-9

Hubbell, E., Liu, W. M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics* 18, 1585–1592. doi: 10.1093/bioinformatics/18.12.1585

Illumina Inc. (2013). *MiSeq QC App*. Available at: http://res.illumina.com/documents/products/appnotes/appnote_miseq_libqc.pdf

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi: 10.1186/gb-2009-10-3-r25

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877. doi: 10.1093/nar/gki901

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi: 10.1186/1471-2105-6-31

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197. doi: 10.1016/0022-2836(81)90087-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 December 2013; accepted: 27 January 2014; published online: 20 February 2014.

Citation: Hadfield J and Eldridge MD (2014) Multi-genome alignment for quality control and contamination screening of next-generation sequencing data. *Front. Genet.* 5:31. doi: 10.3389/fgene.2014.00031

This article was submitted to *Bioinformatics and Computational Biology*, a section of the journal *Frontiers in Genetics*.

Copyright © 2014 Hadfield and Eldridge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.